Genetic Diversity and Its Consequences for Light Adaptation in *Prochlorococcus*

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
AUG 0 3 2011
LIBRARIES
ARCHIVES

by

Gregory C. Kettler

B.A., University of Chicago (2002)

Submitted to the Department of Biology
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Biology

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2011

© Massachusetts Institute of Technology 2011. All rights reserved.

Author .....
Department of Biology
July 25, 2011

Certified by .....
Sallie W. Chisholm
Professor of Biology
Lee and Geraldine Martin Professor of Environmental Studies
Thesis Supervisor

Accepted by .....
Robert T. Sauer
Salvador E. Luria Professor of Biology
Chairperson, Graduate Committee

Genetic Diversity and Its Consequences for Light Adaptation in *Prochlorococcus*

by

Gregory C. Kettler

Submitted to the Department of Biology
on July 25, 2011, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Biology

Abstract

When different cells thrive across diverse environments, their genetic differences can reveal what genes are essential to survival in a particular environment. *Prochlorococcus*, a cyanobacterium that dominates the open ocean, offers an opportunity to explore such differences. Its diversity is examined here, beginning with an overview and comparison of 12 fully sequenced *Prochlorococcus* genomes. The *Prochlorococcus* core genome, that set of genes shared by all cultured *Prochlorococcus*, appears to be well defined by the set shared by these isolates. The flexible genome, that set of genes found in some isolates but not shared by all *Prochlorococcus*, was found to be much larger and open-ended. Most laterally-acquired genes were found to be located in highly variable islands such as those described in previous studies of *Prochlorococcus*. Those lateral gene transfer events can also be placed on the *Prochlorococcus* phylogenetic tree: each *Prochlorococcus* isolate possesses a significant number of genes that even its closest sequenced cousin does not.

A particular gene family may define a *Prochlorococcus* ecotype if those genes are possessed by all members of that ecotype, and if their presence gives that ecotype a selective advantage in some circumstance, thus contributing to the determination of its niche. One gene family is conspicuous for appearing in many copies per genome in one *Prochlorococcus* clade, referred to as eNATL. The sequenced strains belonging to this clade each possess over 40 copies of genes encoding high light inducible proteins (HLIPs), compared to only 9-24 in the other *Prochlorococcus* genomes. Other studies suggest these genes may be involved in resistance to sudden increases in light intensity, among other stresses. This becomes especially interesting as recent field studies also found that eNATL cells may survive changes in light intensity more easily than other low-light adapted *Prochlorococcus*. Here, the effects of light shocks on an eNATL strain and on other *Prochlorococcus* strains are studied. eNATL cultures do recover from light shock conditions that are lethal to other low light-adapted *Prochlorococcus*. Measurements of bulk *in vivo* chlorophyll fluorescence, fluorescence per cell, and variable fluorescence, along with preliminary gene expression data, suggest that the early, rapid response of high light-adapted cells and of eNATL cells distinguish them from other low light-adapted cells, possibly explaining their subsequent survival. The possible role of HLIPs in this response is discussed.

The discussion of HLIPs and eNATL is based on the complete sequences of only two eNATL genomes, both sampled from the same part of the ocean at the same time. That dataset is expanded by the inclusion of Global Ocean Survey environmental shotgun reads, from which are identified several thousand HLIP genes. Past work has shown that HLIPs are divided into two distinct clades: the core, freshwater cyanobacteria-like HLIPs and the flexible, phage-like, island-bound copies. That distinction is examined in the metagenomic data, demonstrating that the separate

types are consistently found in distinct chromosomal neighborhoods. The evolution of HLIPs is also explored by the analysis of large-insert environmental clones containing islands from a variety of eNATL cells. Here, not even all island-bound, HLIP-encoding genes appear to be alike, as only a subset are consistently found in the same locations across the whole eNATL clade.

Ecotype-defining genes are those genes, shared by all members of an ecotype, that provide an ecologically significant advantage, thus helping to define the ecotype's niche. It can be expected that, as environmental data accumulates (including additional measurements of *Prochlorococcus* abundance and newly sequenced genomes from uncultured cells), additional such genes can be identified. This work should represent a model for searching for and examining such genes. Hopefully, future experiments will be able to test the physiological significance of candidate ecotype-defining genes, while feeding back to the environmental data to verify their importance in the open ocean.

Thesis Supervisor: Sallie W. Chisholm
Title: Professor of Biology
Lee and Geraldine Martin Professor of Environmental Studies

Acknowledgments

In the Chisholm lab, I thank Debbie Lindell and Allison Coe for getting me started in the wet lab. I thank Maureen Coleman for suggestions that led to the experiments described here. Her perspective on *Prochlorococcus* evolution became my starting point. I thank the rest of the lab for a vast quantity of feedback over the years. Most of all, I have to thank Penny for her advice and endless encouragement. I could not have finished this work without either. I thank my entire family for their encouragement: Mom, Dad, Sara, and of course David, who I'm sure would agree: this wasn't a race.

This work was supported by a National Institutes of Health training grant, the Department of Energy, the National Science Foundation, and the Gordon and Betty Moore Foundation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The challenge of environmental genomics

One of the central challenges of modern biology is the assignment of particular gene or protein sequences to specific cellular functions. The era of genomics has only accelerated the gene- and sequence-centricity of the field. Nowhere is this more evident than in environmental microbiology. While we have known for some time that there are an unknown number of uncultured, unsequenced microbes in a given volume of soil or water, that invisible mass of cells was, until recently, seen as too intractable to be studied in detail. That is starting to change. In part, this is due to an increasing effort to culture the uncultured microbes (Rappe et al., 2002), but it is especially the increasing power of DNA sequencing that provides the opportunity to study previously unknown microbial populations. Merely sequencing ribosomal DNA gave us a snapshot of the degree of microbial diversity in any given environment (Pace, 1997). But still unclear is the meaning of that diversity *vis a vis* the biogeochemical functions of different microbial groups. The question of how many functionally distinct clades there are within a group of bacteria leads to the question of whether they share one niche, or they each possess closely related niches in the same location. Ribosomal DNA sequencing has since been followed by the shotgun sequencing of the entire metagenome at a given location (Tyson et al., 2004; Tringe et al., 2005; Rusch et al., 2007). This gives us an even greater understanding of the genetic repertoire available within a microbial community, and some idea of its diversity, with the caveat that we often cannot say which variations are functionally significant and which are genetic drift (Wilmes et al., 2008).

The challenge, then: given an increasing knowledge of the genetic repertoire within a microbial community, and some (quite limited) knowledge of the challenges imposed by a local environment,

can we say which genes are responsible for adaptation to which challenge? Sometimes, we can approach that problem through modeling (Bragg et al., 2010), or by watching the evolution of cultures in the unnatural habitat that is the laboratory (Cooper and Lenski, 2000). However, the ideal system in which to answer the above question would be an organism numerous throughout a large but varied environment, while also grown in laboratory cultures where particular hypotheses arising from environmental data might be tested. *Prochlorococcus*, by meeting these criteria, is in some ways the ideal system in which to approach this challenge.

1.2 *Prochlorococcus* as a model system for environmental genomics

Prochlorococcus is a marine cyanobacterium, with the smallest cell size and genome size of any oxygenic phototroph. It is also the most numerous, being found at concentrations on the order of 10^5 cells per mL throughout much of the open ocean (Partensky et al., 1999). It is especially associated with the most nutrient-poor parts of the ocean (Figure 1-1). Due particularly to that scarcity of nutrients, these are difficult habitats for other marine phototrophs such as the larger cyanobacterium, *Synechococcus*, or eukaryotic photosynthesizers. However, their exclusion represents an opportunity for *Prochlorococcus*, which appears to be better adapted to growth in such an environment. Of course, *Prochlorococcus* faces pressures beyond nutrient scarcity: it must cope with predation both by grazers (Jürgens and Matz, 2002; Frias-Lopez et al., 2009) and by phage (Waterbury and Valois, 1993; Sullivan et al., 2003).

For its numbers alone, it would be worth studying *Prochlorococcus*' impact on global ecosystems. It has other attributes, however, that offer insights into the nature and origins of microbial diversity. Most importantly, it is a diverse group. Different *Prochlorococcus* genomes are shaped by different sets of selective pressures, and they have found a variety of strategies to answer those presssures. When we speak of *Prochlorococcus*, we actually speak of a group of closely related (down to 97% similarity in 16S ribosomal RNA sequence), but genetically and phenotypically distinct groups of cyanobacteria, which collectively thrive in a greater variety of environments than one clonal strain of bacteria ever could (Moore et al., 1998). The *Prochlorococcus* phylogeny can be divided into high-light (HL) and low-light (LL) clades, a phenotypic division (on the basis of the ideal light intensity for fastest growth) but also a phylogenetic one: that is, the division of HL and LL ancestors took place only once, early in the history of *Prochlorococcus* (Figure 1-2) (Moore and Chisholm, 1999). It is now known that HL cells are most abundant in the mixed layer close to the surface of the

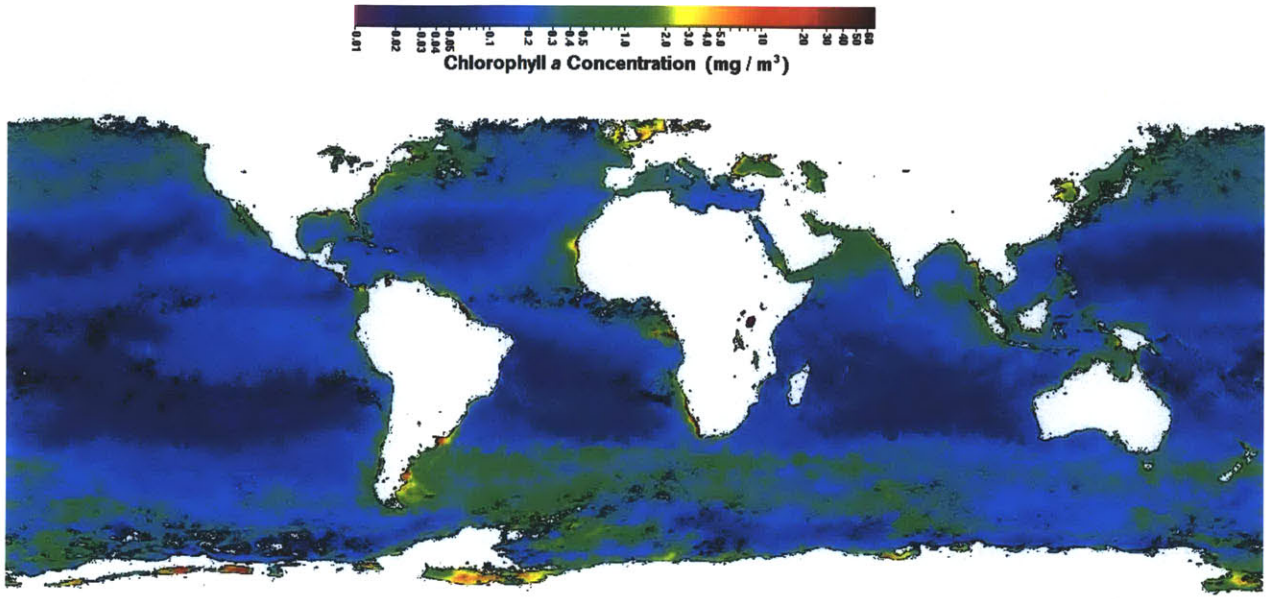


Figure 1-1: Solar-induced chlorophyll fluorescence in the ocean as measured by the NASA SeaWiFS instrument, January 1998. The dark blue regions are the most nutrient-poor; *Prochlorococcus* accounts for up to 50% of the chlorophyll in many of those areas (Partensky et al., 1999).

water column (but they exist in large numbers at lower depths as well), while the LL cells are almost absent from the mixed layer at most times of year, being found primarily at lower depths (West and Scanlan, 1999; Ahlgren et al., 2006). These HL and LL clades can be further divided on the basis of adaptation to different environments on the basis of temperature, nutrient availability, or other factors. The effect of phage predation, for example, is evident in the complements of cell surface-modifying genes that vary greatly even between closely related genomes (Coleman et al., 2006).

1.2.1 *Prochlorococcus* in the wild: the search for selective pressures

Our understanding of this diversity depends on physiological studies of cultured *Prochlorococcus* isolates, which provide an understanding of the environments in which they thrive, and on the sequences of representative genomes from each of these clades, which provide clues as to the mechanisms underlying those physiological properties.

But that alone would not be enough to fully explain the significance of different physiological properties, or of different genes, to a cell growing in the wild. The *Prochlorococcus* ocean environment is a natural laboratory in which to test the effects of various selective parameters. In quantitative PCR (QPCR), we have a means to observe the results of that natural experiment, drawing

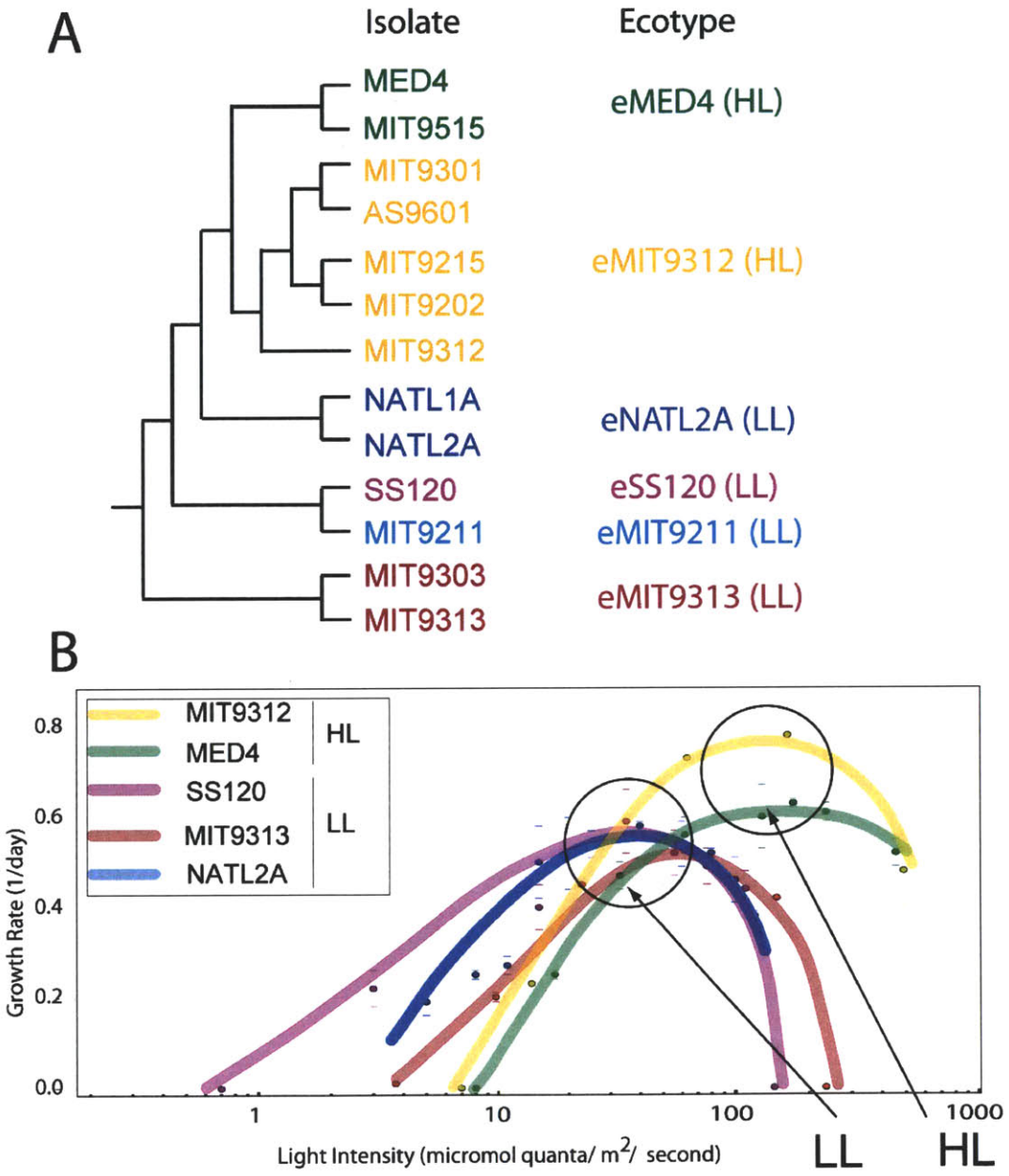


Figure 1-2: (A) Whole genome tree of *Prochlorococcus* as reported in ((Kettler et al., 2007), chapter 2). MIT9202 has been sequenced and added since that publication (Thompson et al., 2011). The number of *hli* genes is as reported in (Coleman and Chisholm, 2007), except for MIT9202 which was analyzed as described in chapter 4. (B) Steady state growth rate of HL and LL *Prochlorococcus* strains as a function of light intensity, replotted from (Moore and Chisholm, 1999) and (Zinser et al., 2007). Circles indicate the optimum light intensity for growth of the LL and HL ecotypes.

direct connections between one environment and one ecotype. QPCR provides accurate counts of each of the major *Prochlorococcus* ecotypes (as defined by their ribosomal inter-transcribed spacer (ITS) sequences) at a given site (Zinser et al., 2006). Comparing their relative abundances at different sites provides insights as to the environmental variables that are most influential in determining their distributions. For example, the abundances measured across a North-South transect of the Atlantic Ocean suggested that temperature determined the respective ranges of two HL-adapted ecotypes, and this hypothesis was substantiated by measurements of their temperature optima in the lab (Johnson et al., 2006). After the earlier division of *Prochlorococcus* into HL and LL-adapted ecotypes (Figure 1-2), this showed further divisions within the HL clade could also be significant. This monophyly of *Prochlorococcus* cells sharing certain ecotype-defining traits (light and temperature) is why QPCR can improve our understanding so well. However, as our attention turns to other properties, such as nutrient scavenging or cell surface modifications, those major clades become less meaningful. This suggests a model in which adaptations to light and temperature took place once, and early in *Prochlorococcus* evolution, but that adaptations to different limiting nutrients, to predation, or to other unknown factors are continuously taking place. Because these adaptations may involve lateral gene transfer, they are not monophyletic (Martiny et al., 2009b).

If that is the case, it necessitates further developments along the path of environmental surveys of *Prochlorococcus* genomes. One approach, now mature but ongoing, is environmental shotgun sequencing (Rusch et al., 2007). These random reads from potentially *Prochlorococcus*-rich waters provide a census of the *Prochlorococcus* metagenome in that area, with comparisons between areas revealing ecologically important genes. For example, where iron or phosphorus are scarce, more genes can be detected encoding their respective starvation-response proteins (Rusch et al., 2010; Coleman and Chisholm, 2010). In the case of iron stress genes, these genes may point to a previously undescribed *Prochlorococcus* clade (Rusch et al., 2010), but the phosphate stress genes may have been recently gained through lateral gene transfer into genomic islands (Coleman and Chisholm, 2010). Such recent events may define even smaller sub-clades, closer to the unexplored leaves of the *Prochlorococcus* phylogenetic tree. Because they are separated by their differing contents of recently acquired genes, these smaller groups may be better understood by longer environmental sequences that show those genes in relation to each other. For example, large-insert fosmid clones allow the comparison of large stretches of different genomes (DeLong et al., 2006), while in the future single-cell sequences will offer the same for entire genomes at a time, without the need to first isolate cultures (Rodrigue et al., 2009).

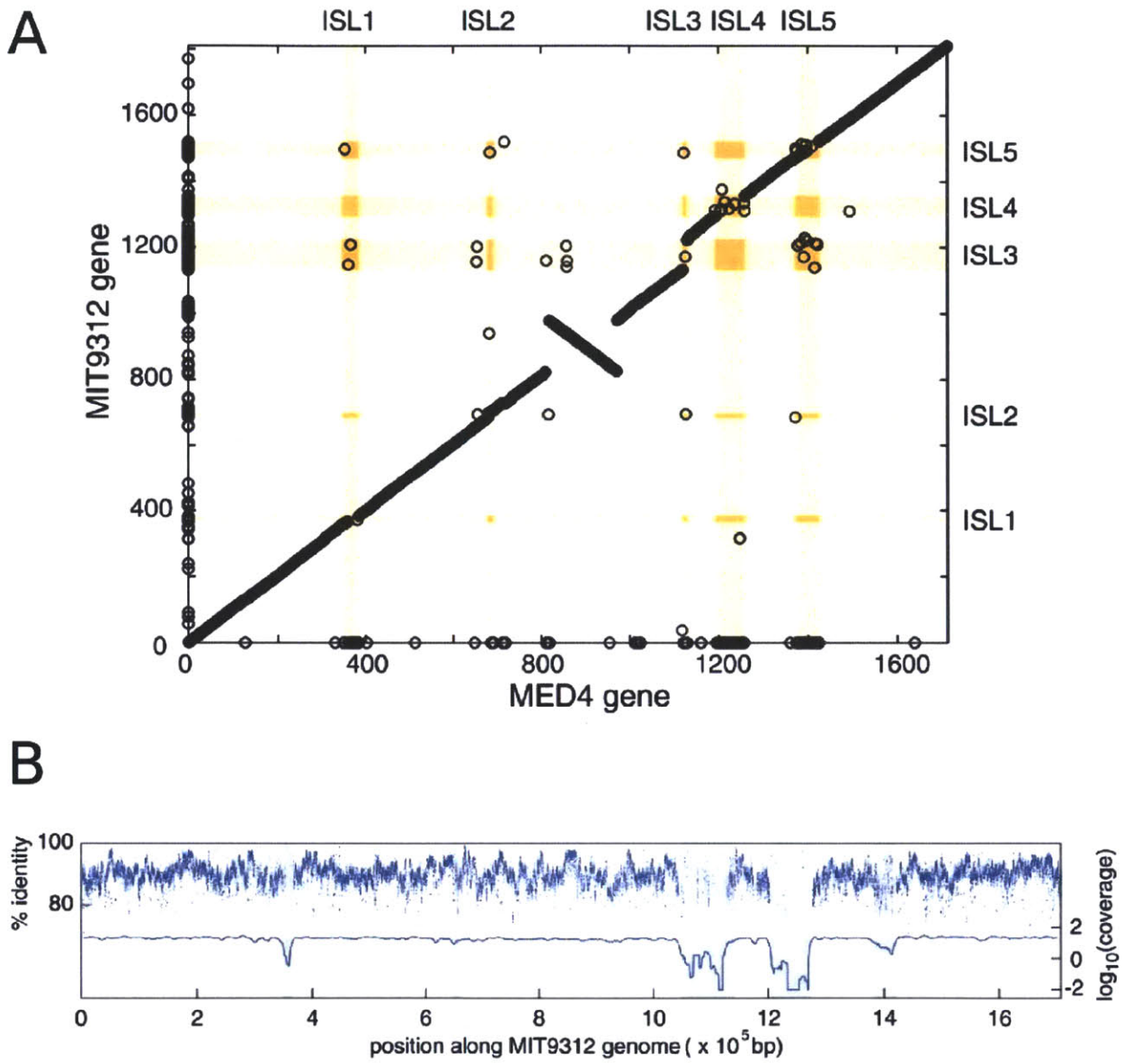


Figure 1-3: Genomic islands of *Prochlorococcus* isolate MIT9312. (A) Synteny analysis between genomes of the MED4 and MIT9312 isolates. Shared genes are plotted with their position on MIT9312 as the y coordinate and their position on MED4 as the x. Isolate-specific genes are plotted on the axes. Shaded regions are islands. (B) Islands as revealed by alignment to randomly sampled environmental DNA fragments (Rusch et al., 2007). The scatterplot reports percent identity of matching reads, and the solid line reports the log of the average number of hits at a given position. Regions of low coverage, indicating greater diversity in that region among the wild population, correspond to islands. Shaded regions correspond to those in (A). (A) and (B) are reprinted from Coleman et al. (2006).

1.2.2 *Prochlorococcus* in the lab: genome to phenotype

The first fully sequenced genomes of *Prochlorococcus* isolates left no doubt that the gene contents of different *Prochlorococcus* isolates had changed dramatically since their divergence, partly by the loss of genes (Rocap et al., 2003; Dufresne et al., 2003). As more genomes became available, analysis of more closely related *Prochlorococcus* isolates revealed that much of the genetic variation was concentrated in a few hypervariable island regions of the chromosome (Figure 1-3) (Coleman et al., 2006). Such islands have been observed in other bacteria, especially pathogens, and the role of phage in transferring those island genes is well established (Mirold et al., 1999; Davis and Waldor, 2003). Phage might also integrate into *Prochlorococcus* genomes as prophages, but this has not been observed in any published *Prochlorococcus* genome. In *Prochlorococcus*, these islands carry a disproportionate number of strain-specific genes, many of them apparently recent acquisitions. Many of those genes are connected to survival in particular habitats, including a those responsible for cell surface variations mentioned above, phosphate metabolism genes, other nutrient transporters, and genes involved in the response to photoinhibition. Further, they contain surprising numbers of cyanophage-like genes, in addition to repeated tRNA sequences that could serve as integration sites between phage DNA and the chromosome (Coleman et al., 2006). Such islands have since been identified in *Synechococcus*. Because fewer closely related *Synechococcus* genomes are available, these were identified on the basis of tetranucleotide frequencies, rather than breaks in synteny (Dufresne et al., 2008).

It was unknown how significant the differences between closely related *Prochlorococcus* isolates might be. In one possible model, adaptations (including lateral gene transfer and genome reduction) took place relatively early in the history of *Prochlorococcus*, leaving relatively minor differences between the most closely related isolates. For example, the adaptation of the eMIT9312 clade to warmer temperatures would be the last major change, separating it from the eMED4 clade. That would leave little difference between the MIT9312 and AS9601 isolates, or between other, to date unsequenced examples of that clade. This can be described as placing most of *Prochlorococcus* genetic variation in the early branches of the *Prochlorococcus* phylogenetic tree. This might simplify the study of *Prochlorococcus* as the existing isolates could be taken as truly representative of their respective clades. The alternative is to place that variation in the leaves of the tree. Here, each *Prochlorococcus* isolate possesses a significant number of genes (or different alleles of the same genes) that even its closest sequenced cousin does not. In this case, while

early adaptation to light and temperature levels would still define the clade as a whole, individual branches within the eMIT9312 clade might possess widely different adaptations to other, as yet undescribed environmental conditions.

The cyanophage, for its part, also incorporates a number of host-like genes and apparently gains a selective advantage from them (Lindell et al., 2004; Sullivan et al., 2005). In particular, phage carry and express genes encoding core photosystem proteins, and they encode genes controlling core carbon metabolism in what may be a mechanism to direct production toward new phage genomes during infection (Lindell et al., 2005; Thompson et al., In press). Whereas cyanophages appear to have acquired their own homologs of host metabolism or stress response genes, these same genes also appear in *Prochlorococcus* genomic islands, suggesting a model in which phage adopt host genes to aid in their own replication, but *Prochlorococcus* cells also re-integrate them into the genome and gain a survival advantage by their presence (Coleman et al., 2006). Such adopted genes include the phosphate stress response genes *pstS* and *phoH*, and plastocyanins and ferredoxins, along with the high light-inducible proteins (HLIPs) discussed below (Sullivan et al., 2005).

While many of these stress response genes might be predicted by their sequences, many other potential stress response genes must exist among the uncharacterized, hypothetical genes. Also, it was initially unknown whether island-located genes would be expressed at all, nor whether two isolates possessing orthologs of the same gene would express them under the same conditions. To address these questions, the whole transcriptome responses to a variety of environmental stresses have been studied in the lab, including starvation for iron, phosphorus, nitrogen (as ammonia or nitrite), and carbon (as carbon dioxide) (Martiny et al., 2006; Tolonen et al., 2006; Thompson et al., 2011; Bagby, 2009). These studies were carried out using the HL MED4 and the LL MIT9313 isolates, using a custom Affymetrix GeneChip that contains probes for both genomes.

Some of these studies have found genes upregulated, and presumably important to the cell's response, that are altogether absent in other strains. For example, one of the most upregulated genes in the MED4 isolate under phosphate starvation, *phoA*, is absent from the MIT9313 isolate (Martiny et al., 2006). Nitrogen starvation, similarly, highlighted possible differences in the regulatory pathways of MED4 and MIT9313, as even shared genes were not always upregulated at the same time when the two isolates were exposed to the same conditions (Tolonen et al., 2006). The same study also demonstrated MIT9313's ability, shared with some but not all *Prochlorococcus*, to use nitrite as a nitrogen source. A timecourse of phage infection demonstrated the importance of the phage's host-like genes, which are expressed early and out of sequence relative to their position

on the phage genome (Lindell et al., 2007, Appendix A). And a series of shifts of MED4 from darkness to white light or filtered light highlighted possible light stress-response genes and suggested the role of a possible blue-light receptor in detecting changes in light level (Steglich et al., 2006).

1.2.3 Differentiating the low light-adapted *Prochlorococcus* clades

Where temperature proved to be a "missing variable" that could separate the two major HL ecotypes (Johnson et al., 2006), we had no such understanding of what might separate the largest LL clades. On the basis of the available genomic and environmental information, tolerance for high light stress now stands out as one possibility. While growth under sustained light intensity was discovered early on to distinguish the two largest clades (Moore and Chisholm, 1999), relatively little attention was paid to short-term increases in light intensity until recently (Six et al., 2007). This became especially intriguing as cells of one LL clade, eNATL, could sometimes be found in greater abundance near the surface than could cells of other LL clades (Zinser et al., 2007).

This suggested a new model to explain their distribution, which follows from the way picoplankton experience mixing in the ocean. During the summer months, wind and waves mix the top of the water column, so a cell in that mixed layer might find itself at the surface or 50 meters deep, both within a period of hours. Not surprisingly, HL-adapted cells dominate here (Zinser et al., 2006; Johnson et al., 2006). This mixed layer has a limited depth, while further down, the ocean is stratified. One can expect a LL-adapted cell and its descendents growing there to remain at that depth through many divisions. In some locations however, during the winter months, the cooling of surface waters leads to convective mixing, effectively extending the mixed layer down to 200 meters or more, for example at the Bermuda Atlantic Timeseries (BATS) site (Steinberg et al., 2001; Lipschultz et al., 2002). In other parts of the world, such as the Hawaii Ocean Timeseries (HOT), winter mixing is less pronounced, so *Prochlorococcus* and other picoplankton populations are more stable at that site year-round (Campbell and Vaulot, 1993). This suggests the possibility that, during a mixing event at the former site, nearly all LL *Prochlorococcus* cells would experience the shock of sudden exposure to high light and, if some had a greater ability to tolerate such a change, they would survive in greater numbers during those mixing events.

1.3 Photosynthesis, photodamage, and photoinhibition

To understand the response of *Prochlorococcus* to mixing, one must consider photodamage in general. All oxygenic phototrophs are subject to some degree of stress thanks to the reducing power generated by sunlight. Photosystem II is particularly vulnerable: one of its core proteins, PsbA (or D1), is damaged by light and must be continuously replaced with a half-life of about 30 minutes (Long et al., 1994; Adir et al., 2003) . However, the precise mechanism of PsbA damage remains controversial. In one model, the manganese-calcium reaction center is damaged directly by light energy (Hakala et al., 2005). In others, reactive oxygen species are generated by photosynthesis (either because plastoquinone pool is fully reduced and cannot accept further electron transfers, or because water oxidation does not proceed quickly enough to re-reduce P680, the primary electron donor) and these react with the PsbA protein (Melis, 1999; Edelman and Mattoo, 2008). Likewise, the role of increasing light in photoinhibition is controversial. It is generally accepted that photoinhibition occurs when the rate of damage to PsbA exceeds the rate of replacement. But this may be due either to an increase in the rate of damage, or to a decrease in the synthesis of PsbA. In the latter model, reactive oxygen species from the photosystem inhibit protein elongation and inhibit the synthesis of any new proteins, including PsbA (Nishiyama et al., 2006). Further past the threshold of photoinhibition, this excess oxidative stress may damage other proteins, leading to a more general stress response and possible cell death (Ziegelhoffer and Donohue, 2009; Mary et al., 2004).

Besides the synthesis of PsbA, photosynthetic organisms have a variety of defenses against these stresses. The simplest is to to capture fewer photons, by adjusting their own light-harvesting complexes (Huner et al., 1998). Absorbed energy must be quenched either photochemically, that is, through productive photosynthesis, or non-photochemically. Non-photochemical quenching (NPQ) generally means "wasting" the energy on a non-productive reaction to release it as heat, for example by the interconversion of xanthophyll pigments (Falkowski and Raven, 2007).

1.3.1 High light inducible proteins

In *Prochlorococcus*, one particular countermeasure against high light stress may play an especially important role: genes encoding the high light-inducible proteins (HLIPs). HLIPs are small proteins (typically less than 70 residues), first described in the cyanobacterium *Synechocystis* sp. PCC7942, and are so named because they were found to be upregulated during times of high light intensity

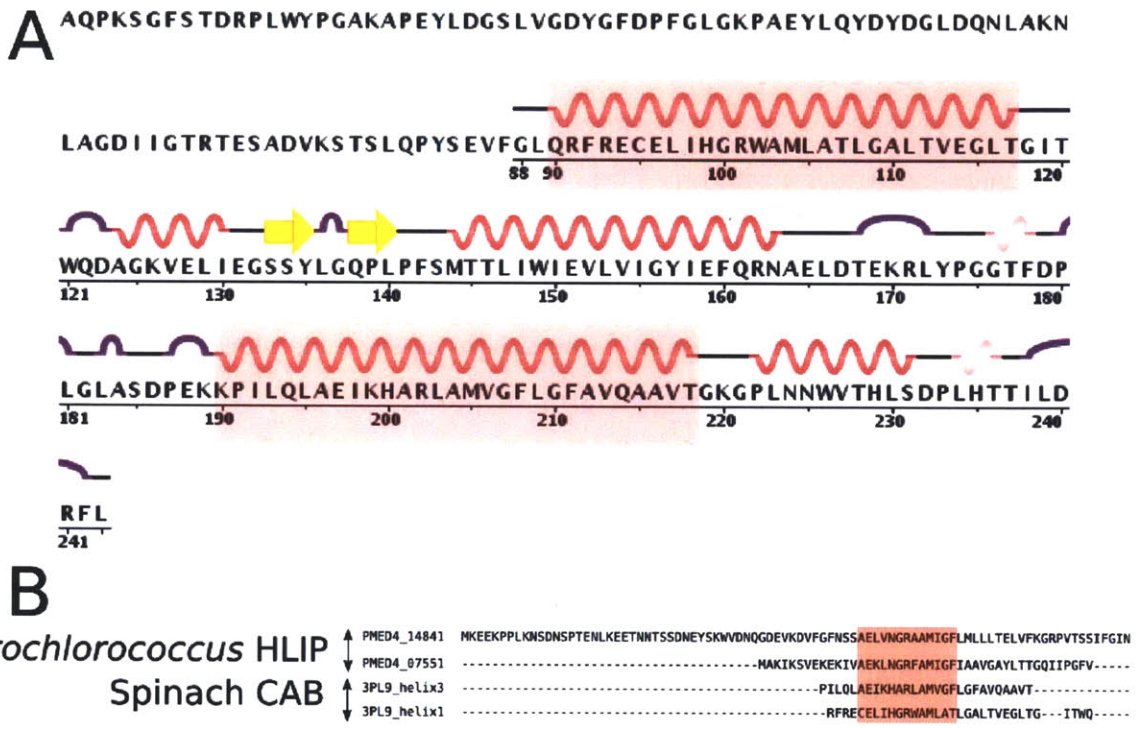


Figure 1-4: HLIPs resemble helices from eukaryotic light-harvesting proteins. (A) The secondary structure of a spinach chlorophyll a/b binding (CAB) protein is depicted. The two highlighted helices have similarity to cyanobacterial HLIPs/SCPs. This CAB structure was reported in Pan et al. (2011), PDB accession 3PL9. (B) Alignment of the two helices from (A) with two *Prochlorococcus* HLIPs.

(Dolganov et al., 1995). Their precise binding partners and mechanism of action is a matter of active research involving *Synechocystis*. However, the first clue to their role was in their sequence, which bears a strong resemblance to the helices of plant chlorophyll a/b-binding (CAB) proteins, part of the light-harvesting complex in eukaryotes (Fig. 1-4). Because they are not always upregulated by high light, but are upregulated by a variety of stress conditions, they have also been called small CAB-like proteins (SCPs), after this resemblance (Funk and Vermaas, 1999).

HLIPs or SCPs are thought to protect the photosystem during periods of stress, but the mechanism by which they do so is unclear. A knockout strain of *Synechocystis* verifies that they are essential to tolerating high light shocks (He et al., 2001). Further investigation of that knockout strain showed it was in some ways hyper-responsive under high light, as it would adjust its pigment complement more than a wild type does. The same investigation found that strain transfers more energy into photochemistry than the wild type did, suggesting that HLIPs play some role in shedding excess light energy as heat (Havaux et al., 2003).

However, the knockout strain was also vulnerable to excess iron, which exacerbates oxidative stress, and knocking out the regulator gene *pfsR* suppresses the lethality of both high light and iron stress, leaving open the possibility of some other role in protecting the cell from oxidative stress (Jantaro et al., 2006). Either way, their precise mechanism is unknown. Studies disagree on which photosystem HLIPs associate with (Yao et al., 2007; Wang et al., 2008). On the basis of their interaction with chlorophyll, it has also been suggested that HLIPs instead protect the photosystem by sequestering chlorophyll during the PsbA replacement cycle (Yao et al., 2007; Vavilin et al., 2007; Nixon et al., 2010).

Synechocystis sp. PCC6803, in which most HLIP studies have since been carried out, has four copies with distinctly different sequences but conserved motifs, plus a fifth, much longer copy that appears to be a fusion of an HLIP with a ferrochelatase domain (He et al., 2001). In *Prochlorococcus*, however, each genome carries between 9 and 40 copies. The upper end of that range occurs only in the low light-adapted eNATL ecotype (Coleman and Chisholm, 2007). If their role in *Prochlorococcus* is indeed connected to a light stress response, this would argue that eNATL is an exception among the LL ecotype. Alternatively, they may be connected to more general stress response, as in MED4 HLIP expression increases during nitrogen starvation (Tolonen et al., 2006), iron starvation (Thompson et al., 2011), and phage infection (Lindell et al., 2007). It is clear HLIPs alone do not explain the entire difference between HL and LL *Prochlorococcus*, as eNATL cells are still best adapted to sustained growth at low light, and even some non-eNATL LL isolates have

almost as many *hli* gene copies as some HL isolates. Instead, their greater copy number in eNATL suggests they are an alternative strategy for light resistance in that one clade.

It is also important to note that in studying the marine cyanobacteria *Prochlorococcus* and *Synechococcus*, the HLIPs are not one, homogeneous family. Genes encoding HLIPs appear in cyanophage as well as cells. These phage *hli* copies then appear to have been re-integrated in host genomes in their highly variable islands (Lindell et al., 2004). Even among the phage-like HLIPs, a variety of sub-families are identifiable, many sharing motifs not seen in *Synechocystis* or any other freshwater HLIP sequences (Bhaya et al., 2002). This leaves open the possibility that some of these sub-families have specific roles besides those being assigned to *Synechocystis* HLIPs.

1.4 Open questions

Connecting genotype, phenotype, and ecotype motivates each of the investigations in this thesis. The first step (chapter 2) was enabled by the sequencing of twelve complete *Prochlorococcus* genomes. Earlier reports of three of those genomes suggested that a process of genome reduction had taken place between the LL- and HL-adapted ecotypes, with the HL-adapted arising more recently (Dufresne et al., 2003; Rocap et al., 2003). A comparison of two closely related genomes within the HL ecotype also pointed to the existence of the islands discussed above Coleman et al. (2006). However, those islands were not described in LL *Prochlorococcus* because no such closely related genomes were available. Furthermore, it was unknown what genes might define the HL and LL ecotypes, or their subclades. Was there a clear HL "signal" in the acquisition of certain genes as the HL genomes branch away from the LL, or are there greater changes within the HL ecotype? The beginnings of a metabolic reconstruction of *Prochlorococcus* were attempted with the SS120 genome (Dufresne et al., 2003), but that may have used genes not shared by other *Prochlorococcus*. It remained to be seen whether other *Prochlorococcus* genomes included all the genes used in that reconstruction. The availability of twelve genomes allows these open questions to be resolved. The approach taken depends on assigning genes to orthologous groups across the twelve genomes, so that shared genes can be identified. This defines the core (shared by all *Prochlorococcus*) genome to be identified, while the pan genome (present in at least one *Prochlorococcus*) is still open. Consideration of shared genes also allows the placement of gene gain and loss events on the *Prochlorococcus* phylogeny, so that potential clade-defining genes might be separated from the variation that exists within each clade. Those same gene gain estimates also inform a new approach to visualizing

Prochlorococcus islands, which is consistent with the results of Coleman et al. (2006) but also can be applied to LL genomes.

Chapter 2 ends with the promise of the hunt for ecotype-defining genes. It leaves open the question of what gene families might be candidates, and how to test them. Building on the results of Zinser et al. (2007), the more recent (Malmstrom et al., 2010, Appendix C) finds that eNATL has a simple phenotype, testable in the lab, with significant consequences for setting it apart from other LL clades in the wild. Chapter 3 further describes that phenotype. Of interest were the behavior of one HL isolate, one eNATL isolate, and two other LL isolates during shocks of a variety of light intensities. The open questions were whether eNATL cells responded to light stress in the same way as true HL cells, and whether HLIPs were involved. Particular attention was paid to the timing of early events. In addition, because a whole-genome expression experiment is still underway, this chapter reports pilot RT-PCR data to demonstrate that the available samples are of significant timepoints in the light stress response.

It is likely, but not certain, that the large number of HLIP-encoding genes in eNATL play a role in this tolerance of changing light levels during mixing events (Coleman and Chisholm, 2007). However, this is based on the complete sequences of only two closely related eNATL genomes, both sampled from the same part of the ocean at the same time. That raises the question of whether the entire eNATL clade has significant numbers of *hli* genes. In addition, past studies, beginning with Lindell et al. (2004), have pointed to the division of *Prochlorococcus* HLIPs into freshwater cyanobacteria-like and phage-like copies. This led us to wonder whether phage-like copies are exclusive to islands. Conversely, we wondered whether the freshwater cyanobacteria-like copies are consistently found in the same genomic neighborhoods, or if they are subject to recombination or lateral gene transfer. These problems can be addressed with the help of metagenomic sequences (large inserts to capture a small island from cells spanning the eNATL clade (DeLong et al., 2006), and shorter shotgun reads from surface waters from the Global Ocean Survey (Rusch et al., 2007). Also, hopefully, other experiments in the lab will be able to test the significance of some of those variations, while feeding back to the environmental data to verify their importance in the open ocean.

Chapter 2

Comparative genomics of cultured *Prochlorococcus* isolates

Patterns and Implications of Gene Gain and Loss in the Evolution of *Prochlorococcus*

Gregory C. Kettler[1,2☯], Adam C. Martiny[2☯¤], Katherine Huang[2], Jeremy Zucker[3], Maureen L. Coleman[2], Sebastien Rodrigue[2], Feng Chen[4], Alla Lapidus[4], Steven Ferriera[5], Justin Johnson[5], Claudia Steglich[6], George M. Church[3], Paul Richardson[4], Sallie W. Chisholm[1,2*]

1 Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, 2 Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, 3 Department of Genetics, Harvard Medical School, Boston, Massachusetts, United States of America, 4 Joint Genome Institute, United States Department of Energy, Walnut Creek, California, United States of America, 5 J. Craig Venter Institute, Rockville, Maryland, United States of America, 6 Department of Biology II/Experimental Bioinformatics, University Freiburg, Freiburg, Germany

***Prochlorococcus* is a marine cyanobacterium that numerically dominates the mid-latitude oceans and is the smallest known oxygenic phototroph. Numerous isolates from diverse areas of the world's oceans have been studied and shown to be physiologically and genetically distinct. All isolates described thus far can be assigned to either a tightly clustered high-light (HL)-adapted clade, or a more divergent low-light (LL)-adapted group. The 16S rRNA sequences of the entire *Prochlorococcus* group differ by at most 3%, and the four initially published genomes revealed patterns of genetic differentiation that help explain physiological differences among the isolates. Here we describe the genomes of eight newly sequenced isolates and combine them with the first four genomes for a comprehensive analysis of the core (shared by all isolates) and flexible genes of the *Prochlorococcus* group, and the patterns of loss and gain of the flexible genes over the course of evolution. There are 1,273 genes that represent the core shared by all 12 genomes. They are apparently sufficient, according to metabolic reconstruction, to encode a functional cell. We describe a phylogeny for all 12 isolates by subjecting their complete proteomes to three different phylogenetic analyses. For each non-core gene, we used a maximum parsimony method to estimate which ancestor likely first acquired or lost each gene. Many of the genetic differences among isolates, especially for genes involved in outer membrane synthesis and nutrient transport, are found within the same clade. Nevertheless, we identified some genes defining HL and LL ecotypes, and clades within these broad ecotypes, helping to demonstrate the basis of HL and LL adaptations in *Prochlorococcus*. Furthermore, our estimates of gene gain events allow us to identify highly variable genomic islands that are not apparent through simple pairwise comparisons. These results emphasize the functional roles, especially those connected to outer membrane synthesis and transport that dominate the flexible genome and set it apart from the core. Besides identifying islands and demonstrating their role throughout the history of *Prochlorococcus*, reconstruction of past gene gains and losses shows that much of the variability exists at the "leaves of the tree," between the most closely related strains. Finally, the identification of core and flexible genes from this 12-genome comparison is largely consistent with the relative frequency of *Prochlorococcus* genes found in global ocean metagenomic databases, further closing the gap between our understanding of these organisms in the lab and the wild.**

Citation: Kettler CG, Martiny AC, Huang K, Zucker J, Coleman ML, et al. (2007) Patterns and Implications of gene gain and loss in the evolution of *Prochlorococcus*. PLoS Genet 3(12): e231. doi:10.1371/journal.pgen.0030231

Introduction

The oceans play a key role in global nutrient cycling and climate regulation. The unicellular cyanobacterium *Prochlorococcus* is an important contributor to these processes, as it accounts for a significant fraction of primary productivity in low- to mid-latitude oceans [1]. *Prochlorococcus* and its close relative, *Synechococcus* [2], are distinguished by their photosynthetic machinery: *Prochlorococcus* uses chlorophyll-binding proteins instead of phycobilisomes for light harvesting and divinyl instead of monovinyl chlorophyll pigments. Although *Prochlorococcus* and *Synechococcus* coexist throughout much of the world's oceans, *Synechococcus* extends into more polar regions and is more abundant in nutrient-rich waters, while *Prochlorococcus* dominates relatively warm, oligotrophic regions and can be found at greater depths [3]. The

Editor: David S. Guttman, University of Toronto, Canada

Received July 30, 2007; **Accepted** November 13, 2007; **Published** December 21, 2007

A previous version of this article appeared as an Early Online Release on November 13, 2007 (doi:10.1371/journal.pgen.0030231.eor).

Copyright: © 2007 Kettler et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abbreviations: HL, high-light; LGT, lateral gene transfer; LL, low-light

* To whom correspondence should be addressed. E-mail: chisholm@mit.edu

☯ These authors contributed equally to this work.

¤ Current address: Department of Earth System Science and Department of Ecology and Evolutionary Biology, University of California, Irvine, California, United States of America

Author Summary

Prochlorococcus—the most abundant photosynthetic microbe living in the vast, nutrient-poor areas of the ocean—is a major contributor to the global carbon cycle. *Prochlorococcus* is composed of closely related, physiologically distinct lineages whose differences enable the group as a whole to proliferate over a broad range of environmental conditions. We compare the genomes of 12 strains of *Prochlorococcus* representing its major lineages in order to identify genetic differences affecting the ecology of different lineages and their evolutionary origin. First, we identify the core genome: the 1,273 genes shared among all strains. This core set of genes encodes the essentials of a functional cell, enabling it to make living matter out of sunlight and carbon dioxide. We then create a genomic tree that maps the gain and loss of non-core genes in individual strains, showing that a striking number of genes are gained or lost even among the most closely related strains. We find that lost and gained genes commonly cluster in highly variable regions called genomic islands. The level of diversity among the non-core genes, and the number of new genes added with each new genome sequenced, suggest far more diversity to be discovered.

Prochlorococcus group consists of two major ecotypes, high-light (HL)-adapted and low-light (LL)-adapted, that are genetically and physiologically distinct [4] and are distributed differently in the water column [5,6]. Given their relatively simple metabolism, well-characterized marine environment, and global abundance, these marine cyanobacteria represent an excellent system for understanding how genetic differences translate to physiological and ecological variation in natural populations.

The first marine cyanobacterial genome sequences suggested progressive genome decay from *Synechococcus* to LL *Prochlorococcus* to HL *Prochlorococcus*, characterized by a reduction in genome size (from 2.4 to 1.7 Mb) and a drop in G + C content from ~59% to ~30% [7–9]. Notably, genes involved in light acclimation and nutrient assimilation appeared to have been sequentially lost, consistent with the niche differentiation observed for these three groups [7]. This comparison suggested that the major clades of marine cyanobacteria differentiated in a stepwise fashion, leading to patterns of gene content that corresponded to the isolates' 16S rRNA phylogeny.

Recently, however, molecular sequence data and physiology studies have revealed complexity beyond the HL/LL paradigms. Within the LL ecotype, for instance, some but not all isolates can use nitrite as a sole nitrogen source [10], and the LL genomes range widely in size [7,8]. Moreover, the distribution of phosphate acquisition genes among *Prochlorococcus* genomes does not correlate to their rRNA phylogeny but instead appears related to phosphate availability: strains isolated from low-phosphate environments are genetically better equipped to deal with phosphate limitation than those from high-phosphate environments, regardless of their 16S rRNA phylogeny [11]. Thus, while the HL/LL distinction has held up both phenotypically and genotypically, there are other differences among isolates that are not consistent with their rRNA phylogeny. Thus, to understand diversification and adaptation in this globally important group, we must characterize the underlying patterns of genome-wide diversity.

Lateral gene transfer (LGT) is one mechanism that creates complex gene distributions and phylogenies incongruent with the rRNA tree. The question of whether a robust organismal phylogeny can be inferred despite extensive LGT is still hotly debated [12,13]. If a core set of genes exists that is resistant to LGT, then gene trees based on these core genes should reflect cell division and vertical descent, as has been argued for the gamma *Proteobacteria* [13]. Others argue that genes in a shared taxon core do not necessarily have the same evolutionary histories, making inference of an organismal phylogeny difficult [14]. In spite of this debate, the core genome remains a useful concept for understanding biological similarity within a taxonomic group. Recent compar-

Table 1. General Characteristics of the *Prochlorococcus* and *Synechococcus* Isolates Used in This Study

Cyanobacterium	Isolate	Light Adaptation	Length (bp)	GC %	Number of Genes[a]	Isolation Depth	Region	Date	Accession Number	Reference
Prochlorococcus	MED4	HL(I)	1,657,990	30.8	1,929	5m	Med. Sea	Jan. 1989	BX548174	[7,38]
	MIT9515[b]	HL(I)	1,704,176	30.8	1,908	15m	Eq. Pacific	Jun. 1995	CP000552	[18]
	MIT9301[b]	HL(II)	1,642,773	31.4	1,907	90m	Sargasso Sea	Jul. 1993	CP000576	[18]
	AS9601[b]	HL(II)	1,669,886	31.3	1,926	50m	Arabian Sea	Nov. 1995	CP000551	[21]
	MIT9215[b]	HL(II)	1,738,790	31.1	1,989	5m	Eq. Pacific	Oct. 1992	CP000825	[19]
	MIT9312	HL(II)	1,709,204	31.2	1,962	135m	Gulf Stream	Jul. 1993	CP000111	[4,60]
	NATL1A[b]	LL(I)	1,864,731	35.1	2,201	30m	N. Atlantic	Apr. 1990	CP000553	[20]
	NATL2A[b]	LL(I)	1,842,899	35	2,158	10m	N. Atlantic	Apr. 1990	CP000095	[22]
	SS120	LL(II)	1,751,080	36.4	1,925	120m	Sargasso Sea	May 1988	AE017126	[8,26]
	MIT9211[b]	LL(III)	1,688,963	38	1,855	83m	Eq. Pacific	Apr. 1992	CP000878	[19]
	MIT9303[b]	LL(IV)	2,682,807	50.1	3,022	100m	Sargasso Sea	Jul. 1992	CP000554	[4]
	MIT9313	LL(IV)	2,410,873	50.7	2,843	135m	Gulf Stream	Jul. 1992	BX548175	[4,27]
Synechococcus	CC9311	Syn.	2,606,748	52.5	3017	95m	Calif. Current	1993	CP000435	[9]
	CC9902	Syn.	2,234,828	54.2	2504	5m	Calif. Current	1999	CP000097	Palenik, unpublished data
	WH8102	Syn.	2,434,428	59.4	2787		Sargasso Sea	Mar. 1981	BX548020	[2,36]
	CC9605	Syn.	2,510,659	59.2	2991	51m	Calif. Current	1996	CP000110	Palenik, unpublished data

[a]Number of protein coding genes excluding pseudogenes

[b]Isolates whose genomes are being reported for the first time here. The gene counts of previously published genomes are slightly different from those of earlier reports [7,8,60] as new annotation pipelines have identified more genes. References refer to either the paper in which the genome was first reported, or the first paper describing the particular isolate.

doi:10.1371/journal.pgen.0030231.t001

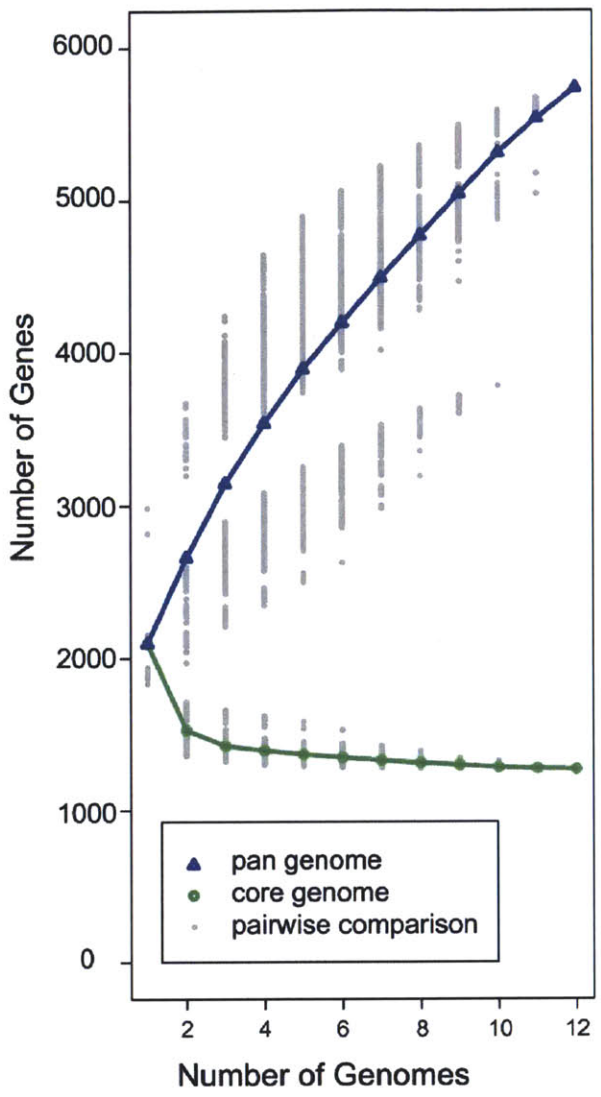


Figure 1. The Sizes of the Core and Pan-Genomes of *Prochlorococcus*
The calculated sizes depend on the number of genomes used in the analysis. If k genomes are selected from 12, there are 12!/(k!(12 − k)!) possible selections from which to calculate the core and pan-genomes. Each possible selection is plotted as a grey point, and the line is drawn through the average. This analysis is based on a similar one in [15].
doi:10.1371/journal.pgen.0030231.g001

isons within the lactic acid bacteria, cyanobacteria, and *Streptococcus agalacticae* groups, for instance, have each revealed a core set of genes shared by all members of the group, on top of which is layered the flexible genome [15–17]. The vast majority of genes in the core genome encode housekeeping functions, while genes in the flexible genome reflect adaptation to specific environments [16] and are often acquired by LGT. Thus the core and flexible genomes are informative not only in a phylogenetic context, for understanding the mechanisms and tempo of genome evolution, but also in an ecological context, for understanding the selective pressures experienced in different environments.

To further understand diversification and adaptation in *Prochlorococcus*, we obtained sequences of eight additional genomes representing diverse lineages, both LL- and HL-adapted, spanning the complete 16S rRNA diversity (97% to 99.93% similarity) of cultured representatives of this group [18–22] (Table 1). Comparing these genomes with available genomes for *Prochlorococcus* and marine *Synechococcus*, our goal was to reconstruct the history of vertical transmission, gene acquisition, and gene loss for these marine cyanobacteria. In particular we identified functions associated with the core and flexible genomes and analyzed the metabolic pathways encoded in each. This analysis reveals not only what differentiates *Synechococcus* from LL *Prochlorococcus* from HL *Prochlorococcus*, but also informs our understanding of how adaptation occurs in the oceans along gradients of light, nutrients, and other environmental factors, providing essential biological context for interpreting rapidly expanding metagenomic datasets.

Results/Discussion

Core Genome

The genomes of 12 *Prochlorococcus* isolates, representing all known major phylogenetic clades, range in size from 1.6 Mbp (MIT9301) to 2.7 Mbp (MIT9303) (Table 1). As more genomes are compared, we observe an asymptotic decline in the number of shared (core) genes (Figure 1), similar to observations for *Streptococcus* genomes [15]. This suggests a finite size of the core genome of approximately 1,250 genes, or 40% to 67% of the genes of any particular isolate. In contrast, the pan-genome [15,23] of these isolates, encompassing the core genes, plus the total of all additional genes found in any of the isolates (the "flexible genes"), contains 5,736 genes (Table S1). The gene accumulation curve as more genomes are added to the analysis is clearly far from saturated (Figure 1), indicating a far larger gene pool within the *Prochlorococcus* clade than is captured by our sequenced isolates, and suggesting the presence of *Prochlorococcus* lineages in the wild, with yet-to-be discovered traits.

Although the closely related marine cyanobacterium *Synechococcus* commonly coexists with *Prochlorococcus*, it is considered more of a generalist, and, collectively, is capable of growth over a broader range of nutrient concentrations and temperatures than is *Prochlorococcus*. To understand the divergence of marine *Synechococcus* and *Prochlorococcus* since their last common ancestor, we looked for genes present in all *Prochlorococcus* but absent from some or all *Synechococcus*. We found 33 such genes, 13 of which are not found in any sequenced marine *Synechococcus* (Table S2). Eight of these *Prochlorococcus*-only genes have been assigned putative functions including one HL inducible protein (MED4's *hli11*, which responds only slightly to light stress [24]), a possible sodium-solute symporter, an iron-sulfur protein, and a *deoR*-like transcription factor, but it is unclear what role these genes have in distinguishing *Prochlorococcus* from *Synechococcus*. Perhaps more importantly, the differentiation between these two groups is defined by the absence in *Prochlorococcus* of 140 genes that are present in all four sequenced marine *Synechococcus* (Table S3). All *Prochlorococcus* isolates sequenced to date lack, for example, divinyl protochlorophyllide a reductase (*dvr*) [25], resulting in one of the defining phenotypic properties of *Prochlorococcus*: divinyl chlorophyll *a* as the primary light harvesting pigment [26]. Other light harvesting genes absent in *Prochlorococcus* include allophyco-

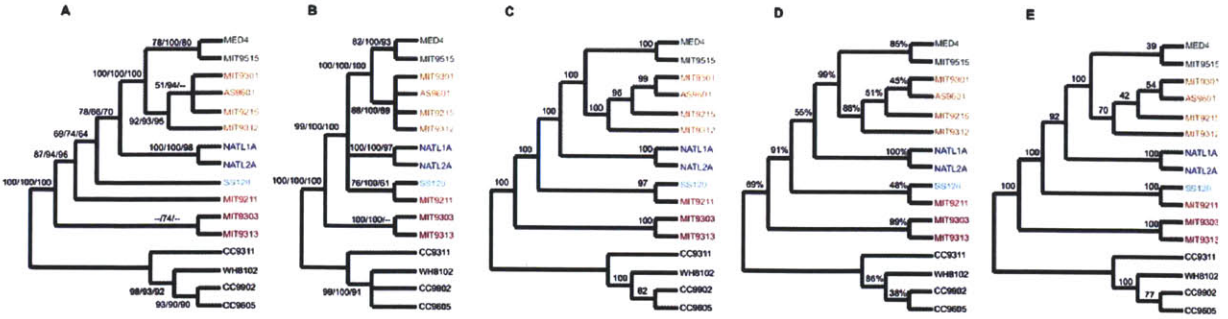


Figure 2. Phylogenetic Relationship of *Prochlorococcus* and *Synechococcus* Reconstructed by Multiple Methods

(A) 16S rRNA and (B) 16S-23S rRNA ITS region reconstructed with maximum parsimony, neighbor-joining, and maximum likelihood. Numbers represent bootstrap values (100 resamplings).

(C) Maximum parsimony reconstruction of random concatenation of 100 protein sequences sampled from core genome. Values represent average bootstrap values (100 resamplings) from 100 random concatenation runs.

(D) Consensus tree of all core genes using maximum parsimony on protein sequence alignments. Values represent fraction of genes supporting each node.

(E) Genome phylogeny based on gene content using the approach of [34]. Values represent bootstrap values from 100 resamplings.

doi:10.1371/journal.pgen.0030231.g002

cyanin (*apcABCDE*), some phycoerythrins, and phycobilisome linkers. *Synechococcus* also possess several molybdopterin biosynthesis enzymes not found in *Prochlorococcus* (*mobA*, *moaABCDE*), which may be necessary for the function of nitrate reductase [27,28]. Although all 12 *Prochlorococcus* isolates also lack the gene for nitrate reductase, this might be a result of the isolation conditions, and further study may reveal nitrate-utilizing isolates [29].

The underpinnings of *Prochlorococcus* diversity should be reflected in the respective roles of the core and flexible genomes. If the core genome provides for central metabolic needs shared by all isolates, it should be possible to reconstruct those pathways with the core genes alone. Therefore we asked whether the core genome encodes all the biochemical pathways needed for growth from the nutrients available to *Prochlorococcus* using Pathway Tools [30] and compared the resulting map with the manually curated, but less detailed, metabolic map for *Prochlorococcus* SS120 [8]. The automated approach is more detailed (Figures S1–S4 and see http://procyc.mit.edu), but the results recapitulate the previous manual effort.

We have identified core genes responsible for nearly all the reactions in the central metabolism, from the Calvin Cycle to the incomplete TCA cycle, including pathways to synthesize all 20 amino acids, several cofactors, and chlorophylls (Figures S1–S4). Among the genes that were assigned functions in the *Prochlorococcus* SS120 core metabolic model, all but seven are found to be part of the core genome in this study. Five of these seven additional genes in SS120 are transporters: SS120_12271, an iron or manganese transporter; SS120_15671, a sodum/alanine symporter; and SS120_06831–06851, three genes encoding an ABC-type amino acid transporter. The other two, *sdhA* and *sdhB*, are putatively responsible for the conversion of fumarate to succinate in the incomplete TCA cycle, but they have no apparent orthologs in many *Prochlorococcus* isolates. Importantly, *sdhAB* in the TCA cycle and *pdxH* in pyridoxal phosphate synthesis are the only cases in which one of the pathways examined could be reconstructed in some strains, but not in the core genome. An additional case, the phosphorylation of pantothenate in coenzyme A synthesis, is incomplete in the core and pan reconstructions, indicating that we have most likely failed to identify the gene or an alternate pathway (Figure S4). This observation supports the view in which essential life functions are unchanging across all *Prochlorococcus*, while nonessential or environment-specific functions are found in the flexible genome (see below). The functions of the latter, then, may relate to niche-specific adaptations that are not required for growth under optimal conditions, but that provide a fitness advantage in particular habitats. The pattern of their gain and loss in phylogenetic space could therefore help us understand when and how *Prochlorococcus* lineages evolved adaptations to particular environments. However, a close examination of their gain and loss requires a robust phylogenetic tree as a scaffold for analysis.

Phylogeny of *Prochlorococcus* Isolates Using the Core Genomes

Identification of the core genome shared by all *Prochlorococcus* isolates provides a new opportunity for determining the phylogenetic relationship among isolates. Our current understanding of the branching order among isolates is based on single gene phylogenies including 16S rRNA [10], 16S-23S rRNA internal transcribed spacer sequence (ITS) [18], *rpoC1* [31], *psbA* [32], and *petBD* [6]. Although trees based on these genes generally agree on the phylogenetic position of most isolates, they disagree, or lack bootstrap support, for the branching order of internal nodes among LL isolates (see Figure 2A and 2B for 16S rRNA and ITS trees). To reconstruct a robust phylogeny, we randomly concatenated 100 protein sequences from a pool of all core genes and compared the topology of the resulting trees (Figure 2C), analogous to the approach described by Rokas and coworkers [33]. This random concatenation was repeated 100 times and the same highly supported topology emerged every time. This tree is very similar to the 16S rRNA tree (Figure 2A) except for the position of LL isolates MIT9211 and SS120. We attribute this discrepancy to the limited information in any single gene (including 16S rRNA), and our analysis suggests that MIT9211 and SS120 form a separate clade. Each node in the concatenated protein tree is also supported by a

plurality of individual core genes (as defined above) (Figure 2D). Based on these results, we postulate that this tree represents the most probable evolutionary relationship among *Prochlorococcus* isolates. However, it is unclear if the physiology of SS120 and MIT9211 warrants considering them as one or separate ecotypes. Furthermore, many single gene phylogenies supported alternative topologies for this node, and future analyses with more genomes or alternative phylogenetic approaches may result in different topologies for this node.

The history of *Prochlorococcus* is marked not only by sequence divergence among the core genes, but also by the gain and loss of genes. We constructed a dendrogram based on the presence or absence of individual orthologous groups (Figure 2E) [34]. Again, the topology of this tree is identical to that of Figure 2C. This suggests that shared gene content among *Prochlorococcus* isolates is significantly influenced by the isolates' phylogenetic relationship despite the occurrence of lateral gene gain and loss.

Flexible Genome

Patterns of gene gain and loss in the evolutionary tree. We used our most probable phylogenetic tree (Figure 2C) as a map for the evolution of each isolate and superimposed the gain and loss of flexible genes (i.e., non-core) upon it (Figure 3A). By assigning costs to gain and loss events (see Methods) and then minimizing the total cost (maximum parsimony criterion), we estimated for each gene in each node of the tree whether it was more likely to have been inherited from a common ancestor or acquired at that node [35].

As mentioned above, 140 genes found in all *Synechococcus* are absent in all *Prochlorococcus* (Table S3). This is consistent with our earlier image, based on only four genomes, of progressive gene loss from *Synechococcus* to LL *Prochlorococcus* to HL *Prochlorococcus* [7,8,36]. However, our analysis suggests an alternative to this view, in that the MIT9313 lineage (i.e., the MIT9313/MIT9303 "cluster" or eMIT9313 clade, *sensu* [37]) is not simply an intermediate step in this gene loss process. Although the genome sizes within eMIT9313 are similar to those of *Synechococcus*, the eMIT9313 clade appears to have gained a large number of genes, including many unique to each isolate. These genes are not found in any other sequenced *Prochlorococcus* or *Synechococcus* strain, and the eMIT9313 strains may therefore have acquired them after their divergence from the other *Prochlorococcus*. The large difference between strains MIT9313 and MIT9303 is then most likely the result of further gene gains after they diverged from each other. After the divergence of eMIT9313, all *Prochlorococcus* genomes have a roughly constant size (1.66 to 1.84 Mbp). However, we still observe significant gene gain and loss. A few particular examples are discussed below, but additional work remains to show how these dynamics contribute to the distribution patterns we observe in the oceans for specific lineages.

Ecotypic differences: Genes underlying the HL/LL ecotypes. As described in many previous studies, *Prochlorococcus* can be classified into two broad groups based on their growth adaptation to specific light intensity (and corresponding phylogeny) [4]. In addition to the core genome shared by all 12 *Prochlorococcus* examined in this study, HL isolates all share an additional 257 genes, 95 of which are not found in any of the LL isolates (Table S6). This HL core provides further clues to the genetic bases for the HL/LL physiological and ecological differentiation that has been observed in previous works [4,5,19,20,37–42]. All HL isolates carry an operon containing a DNA ligase, exonuclease, and helicase, which might be involved in DNA repair or other nucleic acid processing. HL isolates also possess large numbers of HLIPs (although NATL1A and NATL2A have more), which are thought to protect photosystems from oxidative damage [39] and are upregulated in stress conditions such as high light [24], nitrogen starvation [43], and phage infection [44]. In particular, they share at least three additional genes for HL inducible proteins not found in any other strain. In addition to HL stress, one (*hli8/18* in MED4) is upregulated in response to phage infection, and the other two (*hli15* and *hli22*) by nitrogen starvation [43,44]. The HL isolates also share some genes with no clear connection to photobiology, such as a uridine kinase that may provide an alternative pathway for uracil recycling to UMP. In all *Prochlorococcus*, UMP can be generated by core pathways involving the core *upp* or *pyrBCDEF* genes [45]. All HL isolates also share the operon *tenA-thiD*, which may be involved in thiamine salvage and/or degradation [46,47]. In addition, the HL core contains dozens of hypothetical and conserved hypothetical genes not found in any LL isolate, and these might be critical for survival in the commonly nutrient-poor, HL environment of the surface oceans. Finally, all HL and eNATL2A isolates (which are LL, but closest to the HL clade) include at least one photolyase (orthologs of P9301_3091) and a second possible (P9301_03091), and some HL strains have a third (P9301_03921), the function of which is to repair UV-induced DNA lesions (Table 2).

Likewise, LL isolates share an additional 92 genes beyond the *Prochlorococcus* core, 48 of which are not found in any HL isolates (Table S7). All *Prochlorococcus* have lost the majority of genes involved in phycobilisome synthesis but LL isolates retain several phycoerythrin genes (*cpeABSTYZ*), whereas HL isolates have lost all but *cpeB* and *cpeS*, consistent with previous observations based on fewer genomes [48]. The role of phycoerythrin in *Prochlorococcus* remains uncertain, but may be related to signal transduction rather than light harvesting [49,50]. Individual *Prochlorococcus* strains possess different complements of amino acid transporters. But all LL isolates, and only some HL isolates, contain the tandemly arranged amino acid transporter components *glnQ* and *hisM*, suggesting some variation among *Prochlorococcus* ecotypes in the ability to take up amino acids [51].

Several exonucleases that repair UV-induced lesions, encoded by *recJ* and *xseA*, are exclusive to LL isolates, which is surprising given their reduced exposure to UV radiation. These genes might be necessary to protect against UV exposure during mixing events, and their absence from HL isolates suggests the HL isolates have different strategies to limit DNA damage. Moreover, LL isolates exclusively encode *mutY*, whose product prevents mutations arising from oxidatively damaged guanine residues [52]. The absence of the *mutY* gene in HL *Prochlorococcus* has been hypothesized to underlie their extremely low %G + C content, by increasing the frequency of G-C to A-T mutations [7]. However, this gene is present in LL isolates with %G + C as low as 35%, suggesting that *mutY* alone is not responsible for genomic A + T enrichment [53].

Ecotypic differences: Clades within the HL and LL

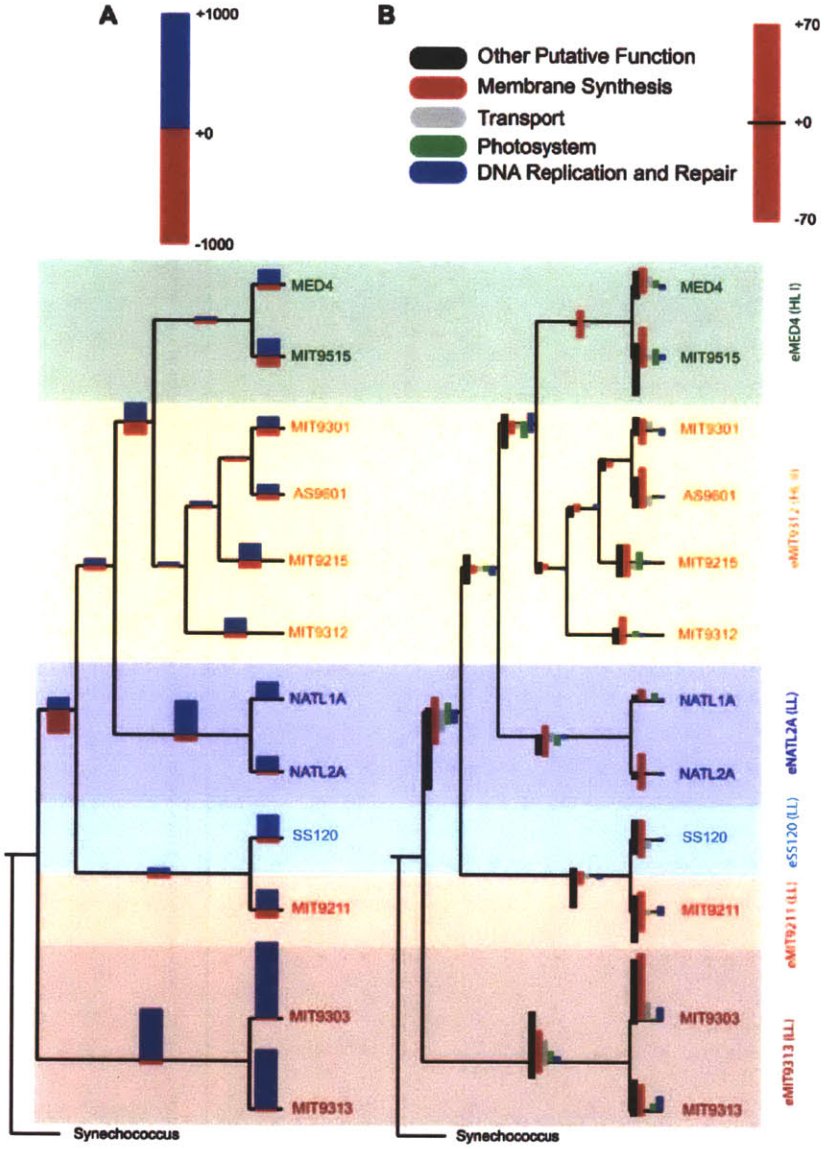


Figure 3. The Loss and Gain of Genes through the Evolution of *Prochlorococcus*

The ancestor node in which a gain or loss event took place was estimated by maximum parsimony. Four marine *Synechococcus* genomes (not shown) were included in the calculation, and the phylogenetic tree from Figure 2C was rooted between the *Synechococcus* and *Prochlorococcus* lineages.
(A) The total number of genes gained and lost at each node.
(B) The loss and gain of genes in that could be assigned functional roles through homology. Note that (B) focuses on the small minority of genes that do have an assigned function. Genes were assigned to one of five categories on the basis of keyword matches against the gene name or COG description. "Other Putative Function" refers to genes with assigned function but not belonging to the four major categories. Note the difference in scale for (A and B).

doi:10.1371/journal.pgen.0030231.g003

ecotypes. Going beyond the HL and LL ecotypes, two distinct subclades have been identified within the HL ecotype (eMED4 and eMIT9312), and several lineages within the LL ecotype (eNATL2A, eMIT9313, and eSS120 + eMIT9211) [18] (Figure 3). The distribution of cells belonging to these subclades has been measured along extensive environmental gradients in the oceans, and the two HL subclades have distinct distributions most strongly correlated with surface temperature [39,40]. Moreover, two LL clades (eNATL2A and eMIT9313) have distinct distributions as well: cells related to eNATL2A can be abundant at the surface, while cells related to eMIT9313 are generally found at the base of the euphotic zone in stratified waters and never at the surface [40]. This is in spite of the two clades' similar optimum light intensity for growth [19,40]. Given these ecological distinctions, we looked for genes distinguishing these subclades (Table 2).

The eMIT9313 clade has many features that distinguish it

Table 2. Non-core Genes Referred to in the Discussion

Section	Gene	Function	Locus	MIT9313	MIT9303	MIT9211	SS120	NATL1A	NATL2A	MIT9301	AS9601	MIT9215	MIT9312	MIT9515	MED4
Underlying the HL/LL ecotypes	*lhr*	Helicase	PMED4_08051							x	x	x	x	x	x
		DNA ligase	PMED4_08061							x	x	x	x	x	x
		Beta-lactamase fold exonuclease	PMED4_08071							x	x	x	x	x	x
	hli8/18	Photosystem protection	PMED4_15941							x	x	x	x	x	x
	hli15	Photosystem protection	PMED4_12761							x	x	x	x	x	x
	hli22	Photosystem protection	PMED4_07541							x	x	x	x	x	x
	udk	Uridine kinase	PMED4_11091					x	x	x	x	x	x	x	x
	tenA-thiD	Thiamine salvage	PMED4_03811–21							x	x	x	x	x	x
	phrB	Photolyase	PMED4_02901					x	x	x	x	x	x	x	x
	phrB	Photolyase	P9301_03921							x	x	x		x	
		Possible photolyase	P9301_04471					x	x	x	x	x	x	x	x
	cpeADR-TYZ	Phycorethryin	non-adjacent	x	x	x	x	x	x						
	cpeBS	Phycorethryin	non-adjacent	x	x	x	x	x	x	x	x	x	x	x	x
	hisM	Amino acid transport	P9313_10601	x	x	x	x	x	x				x	x	
	glnQ	Amino acid transport	P9313_10611	x	x	x	x	x	x	x			x	x	
	recJ	Exonuclease	P9313_08931	x	x	x	x	x	x						
	xseA	Exonuclease	P9313_20731	x	x	x	x	x	x						
	mutY	Mismatch repair	P9313_01441	x	x	x	x	x	x						
Within the HL/LL ecotypes		Sigma factor	P9313_13631	x	x										
		Sigma factor	P9313_27801	x	x										
		Sigma factor	A9601_12341								x				
		Sigma factor	P9313_09171	x	x										
	gdhA (1)	Amino acid synthesis	P9313_07431	x	x										
	gdhA (2)	Amino acid synthesis	P9515_04091									x		x	
	cytA	Electron transporter	P9313_06071	x	x										
	cypX	Electron transporter	P9313_19741	x	x										
	melB	Disaccharide transport	P9211_03411			x	x								
	glcD	Dehydrogenase	P9211_13031			x	x								
	citB-baeS	Signal Transduction	P9211_15001–11			x	x								
		Disulfide bond formation	P9211_15411			x	x								
	nirA	Nitrite reductase	P9313_28061	x	x			x	x						
	sdhA	Possibly TCA cycle	A9601_12591							x	x	x	x		
	sdhAB	Possibly TCA cycle	P9313_01411–21	x	x		x								
	phoBR		PMED4_07791–801	x[a]	x			x	x	x			x		x
	phoE		PMED4_07831	x	x			x	x	x			x		x
	cynS	Cyanate lyase	PMED4_04061					x	x						x
	amtB (1)	Ammonia permease	PMED4_02681	x	x	x	x	x	x	x	x	x	x	x	x
	amtB (2)	Ammonia permease	P9515_04231											x	
	urtBCD	Urea transport	PMED4_10831–51	x	x			x	x	x	x	x	x		x

Each line is an orthologous group, for which the gene name and putative function are given, if available. The locus given is that of an arbitrarily selected gene in the group; the complete list for any orthologous group is available in Table S3. The presence or absence in each *Prochlorococcus* isolate is given [11].

[a]*phoR* is not functional in MIT9313.

doi:10.1371/journal.pgen.0030231.t002

from other *Prochlorococcus* (Table S8). Acquired genes include multiple sigma factors and kinases, likely involved in signal transduction, outer membrane synthesis enzymes, and transporters. Their possession of transporters not found in other *Prochlorococcus* or in *Synechococcus* may imply that they are exploiting nutrient resources unique to their environment, or they may simply have experienced weaker selection for reducing genome size. Likewise, the two isolates in this clade (MIT9313 and MIT9303) share three sigma factors (MIT9303 has a fourth) and several other transcriptional regulators not found in any other isolate, suggesting they have more complexity in their ability to respond to various stimuli. The eMIT9313 isolates also share a glutamate dehydrogenase gene (*gdhA*), absent from most other *Prochlorococcus* (two HL isolates share a distantly related allele), which provides an alternative pathway for ammonium incorporation besides the standard GS-GOGAT pathway. This enzyme has been shown in *Synechocystis* to be important during the late stages of growth when energy is limiting, and for ammonia detoxification [54]. We also observe that photosystem II genes *psbU* and *psbV* are exclusively found in eMIT9313 (as well as most other cyanobacteria) along with possible electron transporters (*cytA*, *cypX*). The eMIT9313 isolates carry only three *pcb* genes, encoding light harvesting antenna proteins, compared to six or seven in the other LL isolates. This relative lack of *pcb* genes, however, does not seem to prevent growth at very low irradiances, as eMIT9313 cells are often found at the base of the euphotic zone. The eMIT9313 isolates also have relatively few genes for HLIPs (nine in eMIT9313, compared to 12–13 in SS120/MIT9211 and 41 in eNATL2A), which might help explain why this clade is not found in surface waters.

Five genes with assigned functions were unique to eSS120/eMIT9211 (P9211_03411, P9211_13031, P9211_15001, P9211_15011, P9211_15411), but there were no clear linkages between these genes and the distribution pattern of this group in the ocean.

In contrast, the eNATL2A isolates (NATL1A and NATL2A), whose low optimum light intensity for growth marks them as LL [19,40] have some notable HL-like properties. The eNATL2A isolates possess photolyase genes, like HL isolates, and they harbor more genes for HLIPs than any other HL or LL isolate. Together these genes may help explain the abundance of eNATL2A at the surface relative to other LL clades [40]. They also share the uridine kinase found in HL isolates.

All isolates in the eMIT9313 and eNATL2A clades possess a nitrite reductase gene, *nirA*, whereas no other *Prochlorococcus* lineages (HL or LL) have this gene, a difference that has been confirmed through physiology studies [10]. The availability of nitrite may therefore influence the distribution of these two clades, although this pattern has not emerged in the field studies to date [39,41].

In spite of their different distributions in the ocean, we could identify only one gene with a described function that distinguishes the two HL clades eMIT9312 and eMED4. All isolates in eMIT9312 possess a gene similar to *sdhA* which encodes succinate dehydrogenase. Unlike the proteobacteria-like *sdhA* found in SS120, MIT9313, and MIT9303 and previously assigned to the incomplete TCA cycle [8], the HL gene is actinobacteria-like and is not accompanied by *sdhB*, raising the possibility that this dehydrogenase/reductase acts on a different substrate. Temperature variability is most strongly correlated with differences in the abundances of eMED4 and eMIT9312 along a longitudinal gradient in the oceans, and this is consistent with the temperature limits for growth for strains representing these ecotypes in culture [39]. These properties could emerge from differences within orthologous proteins, yielding different enzymatic reaction temperature optima, rather than from the presence or absence of entire genes. This complicates the search for ecotype-defining genes in their case.

Isolate-specific genes. We found that a large fraction of variability was in the "leaves of the tree," that is, genes gained by one isolate but not necessarily by others in the same clade (Figure 3B and Table S4). The greatest differentiator between the most closely related isolates are genes related to outer membrane synthesis (Table S5). For example, while MIT9515 and MED4 each have several genes in COG438 and COG451 (both COGs described as acyltransferases connected to outer membrane synthesis), these genes are only distantly related [55]. Six genes matching COG438 are found in MIT9515 but not MED4, and these six all have best matches to genes in lineages outside *Prochlorococcus*. The rapid turnover of genomic content contrasts with the broader similarity of their roles: even though the genes found in different isolates are not orthologs and have little to no sequence similarity, they share the same biological role. Such membrane synthesis genes were probably lost or gained continuously throughout the evolution of *Prochlorococcus*, as every ancestor node is estimated to have lost or gained some in that category (Figure 3B).

Certain cell surface proteins are potentially under strongly diversifying selection if they serve as attachment or recognition sites for predators or phages. The observed variation among genomes in relation to this category supports this idea and suggests that the predatory environment could be different in each of the locations where these isolates originated. However, it is deceptive to consider these the most recent changes, as there are innumerable undiscovered *Prochlorococcus* genotypes in the wild, some of which could fill the gap between MIT9515 and MED4, for example. Such variation, some of which may be adaptive, is below the resolution of current methods for measuring ecotype abundance in the oceans [39,42,56].

After cell surface synthesis, the next largest fraction of the flexible genome is transporters (Figure 3B). As discussed above, the larger genomes of MIT9303 and MIT9313 have a significant number of transporters not shared with other *Prochlorococcus*, although some are shared with *Synechococcus*. Among their predicted substrates are toxins, sugars, and metal ions. Relatively few transporters are specific to the other LL isolates. In addition, each HL isolate possesses a different set of transporters, but there is no set both universal among HL isolates and absent from LL isolates. Furthermore, the presence of specific transporters does not follow the phylogeny of the HL ecotype. Transport genes must therefore be subject to rapid gain and loss, such that their presence is not conserved within the subclades. Transport reactions are peripheral to metabolic pathways, and such peripheral reactions are predicted to be subject to the most rapid turnover [57].

Individual *Prochlorococcus* isolates also contain multiple copies of specific light-related genes but in different

numbers. MED4, the first HL genome to be studied, has only one *pcb* light harvesting antenna gene whereas the first LL genomes had two (MIT9313) or eight (SS120) [58]. Our new data identify MED4 as the exception, since the other five HL isolates share a second copy in the same well-conserved neighborhood. Surprisingly, there is huge variation in the number of genes encoding HLIPs, ranging from nine in eMIT9313 to 41 in eNATL2A. Even at the leaves of the tree, within the HL clades, HLIPs range in copy number from 15 to 24.

A second copy of the core photosystem II gene *psbA* also appears in more than half the genomes. This gene is especially interesting because it is also found in all *Prochlorococcus*-infecting myoviruses and podoviruses sequenced to date [59]. While it is possible that *psbA* might have been inserted into the genome by those viruses, much as the genes in genomic islands are thought to have been [60], the similarity between *psbA* copies in the same genome suggests they are the result of intragenomic duplication events, not transduction. Indeed, in all of these strains the two copies are identical or nearly identical in nucleotide sequence, suggesting that they result from a very recent duplication event. Furthermore, while extra *psbA* copies sometimes appear in islands, they do not always. In MIT9515, for example, the two copies lie in tandem but not in an island. It is not clear why *psbA* is subject to such duplication events while other photosystem genes are not. The most likely reason is that the PsbA protein (D1) has an exceptionally brief half-life due to light-induced damage [61], and therefore two gene copies help ensure sufficient product via a gene dosage effect and/or by promoter differences leading to expression under different conditions.

The complement of nutrient assimilation genes also varies among the most closely related isolates, suggesting frequent gain and loss events. Such variability was recently described for genes involved in phosphorus assimilation [11]. Within the eMIT9312 clade, for instance, the isolates AS9601 and MIT9215 are lacking the *phoBR* two-component system, the *phoE* porin, and several related genes that are present in MIT9312 and MIT9301. Now equipped with whole genomes for 12 isolates, we see a similar situation for nitrogen assimilation genes. MED4 is the only HL isolate with cyanate lyase, and likewise MIT9515 exclusively carries a second ammonia permease gene. In contrast, MIT9515 is the only HL isolate lacking urea transport and metabolism genes. This variability may reflect the available nitrogen sources in the local environment where these isolates originated, as has been hypothesized for phosphorus [11].

Chromosomal Location of the Flexible Genome

Previous work comparing the genomes of two closely related *Prochlorococcus* isolates has highlighted the importance of highly variable island regions in genomes as the sites of genomic variation [60]. These variable genome segments appear to contain genes that could be important for adaptation to local conditions, and include many of the functions encoded in the flexible genome analyzed here, such as outer membrane synthesis. Thus, we analyzed the chromosomal geography of the flexible genome. Are flexible genes preferentially located in island regions, and if so are the most recently acquired genes more likely to be island genes?

To answer these questions, we plotted the timing of gene gain events against their chromosome positions (Figures 4 and S5 and S6). In HL isolates, the islands contain the majority of gained genes. Furthermore, the islands include not only recent acquisitions but also genes that were gained long ago, based on their presence in divergent modern isolates. However, particular islands show different levels of gain or loss events throughout the evolution of *Prochlorococcus*. Apparently, these sites have been important for adaptation throughout the history of most *Prochlorococcus* lineages.

In the earlier comparison of two genomes at a time, islands were identified as breaks in syntenic regions [60]. Among LL isolates, this approach is difficult because the genomes are more divergent, and numerous rearrangements have disrupted synteny, even for core genes. Plotting gene gain events along the chromosome, however, reveals island structure in several LL genomes. MIT9211 and SS120 have clearly defined islands much like the HL isolates, while NATL1A and NATL2A have one large potential island and several much smaller sites (Figures 4 and S5 and S6).

Surprisingly, this approach is less helpful in the two large genomes, MIT9313 and MIT9303, which have apparently gained a large number of genes throughout the chromosome (Figure 4). In their organization and content, the large genomes are exceptional among *Prochlorococcus* in three ways: they share a large number of genes with *Synechococcus* that the other isolates do not, they gain additional genes not shared with any other *Prochlorococcus* or with marine *Synechococcus*, and those genes do not cluster in discernible islands. The first two differences mean that their genome sizes are much greater than those of the other isolates. The lack of islands together with the larger genome size could indicate that these isolates have acquired genes through a different mechanism that does not direct them toward islands. The relative lack of pressure towards genome reduction in the evolution of eMIT9313 may also play a role. However, additional sequenced genomes may provide better coverage of the eMIT9313 clade and clarify the timing of gene gain events.

The Frequency of Core and Flexible Genes in Wild Populations

Because *Prochlorococcus* is very abundant in many regions of the oceans that have recently been sampled and subjected to metagenomic analysis [62–64], we have an opportunity to test the robustness of our distinction between core and flexible genes in *Prochlorococcus*. If the core genome we have defined, based on the genomes of 12 isolates, is reasonably universal and core genes are generally single copy per genome, we would expect to find core genes represented with equal frequency in the ocean; the occurrence of non-core genes, in contrast, would be more variable. To test this hypothesis, we used the MIT9301 core and flexible genomes as queries against the Global Ocean Survey dataset [64], as MIT9301 often shares the highest sequence similarity with GOS sequences. As expected, the core genes, after normalization to gene size, are represented in roughly equal abundance in the database, with only a few exceptions (Figure 5A). In the case of non-core, or flexible genes, many had few or no hits, and a few were even more abundant than the average core gene, suggesting more than one copy per genome (Figure 5A). Seven core genes are underrepresented in the GOS dataset relative to other core genes, and all seven are located in a genomic island in MIT9301 largely related to cell surface

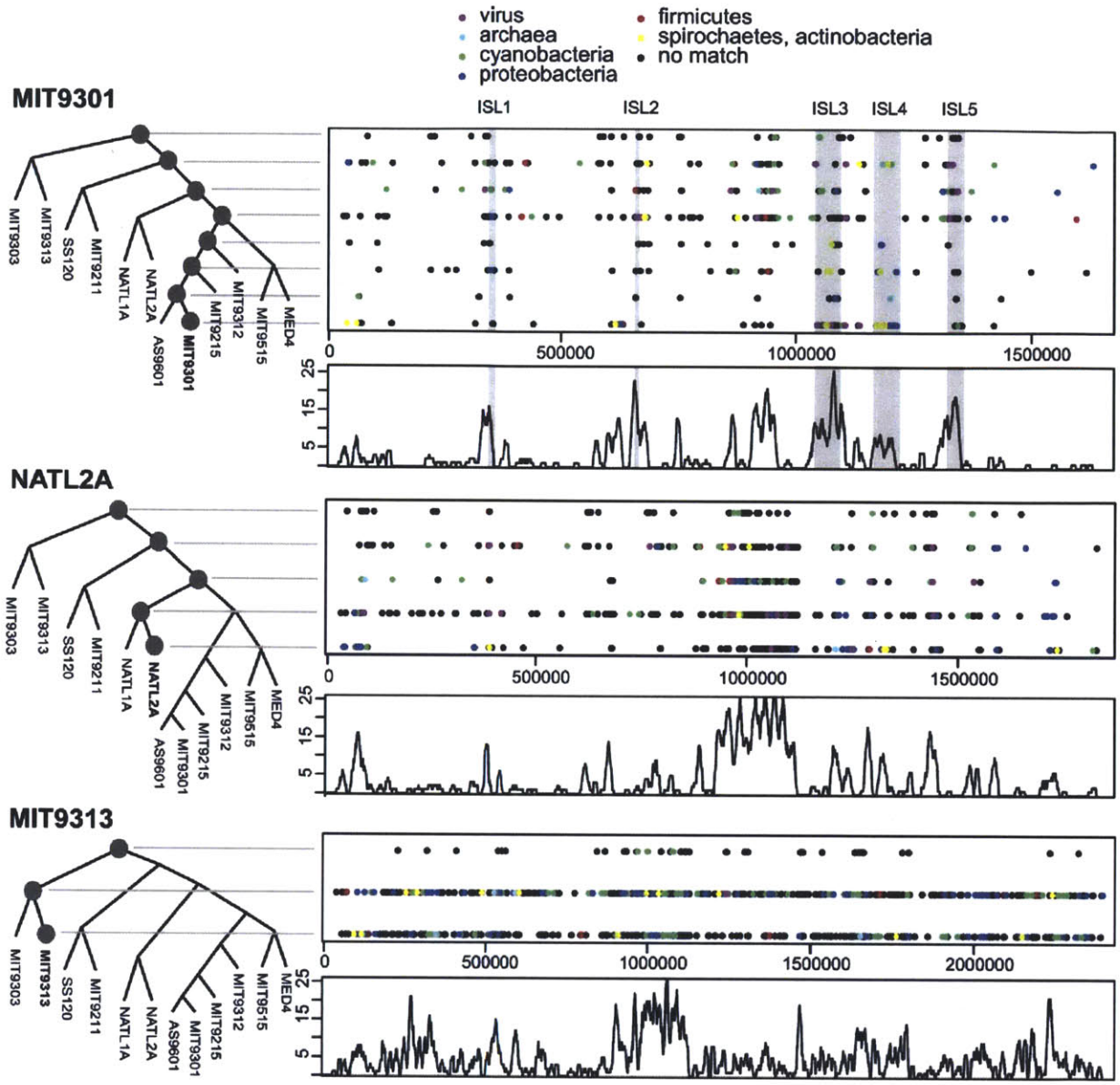


Figure 4. Gene Acquisitions Confirm Known, and Identify Novel, Genomic Islands in *Prochlorococcus*

The dot plots indicate the location on the chromosome and the ancestor node in which the gene is estimated to be gained. The color indicates where the best match was found. In MIT9301, The shaded regions are islands as defined by [60]. Gained genes are defined for each node as in Figure 3. The lower plot is the number of genes gained in a sliding window (size 10,000 bp, interval 1,000 bp) along the chromosome.

doi:10.1371/journal.pgen.0030231.g004

biosynthesis (Figure 5B). The most abundant flexible genes encode HLIPs and hypothetical proteins and are also found in islands in MIT9301 (Figure 5B). This supports the hypothesis that islands are dynamic reservoirs for recent and local adaptation.

Conclusion

In this study we have attempted to advance our understanding of the evolutionary origins of diversity in *Prochlorococcus* by defining the core and flexible genomes and examining the patterns of gain and loss of non-core genes over the course of evolution. We have learned, for example, that many genes involved in adaptation to different light intensities and DNA repair were apparently fixed before the modern clades diverged, and as a result, the HL-/LL-adapted dichotomy has persisted both genetically and phenotypically. The eNATL2A clade appears to be a refinement on the HL/LL paradigm, as its isolates grow optimally at light intensities typical of the LL ecotype, but have the photoprotective abilities of the HL ecotype. More recent changes in genome content, i.e., those occurring at the tips of the phylogenetic tree, involve cell surface features that are likely under

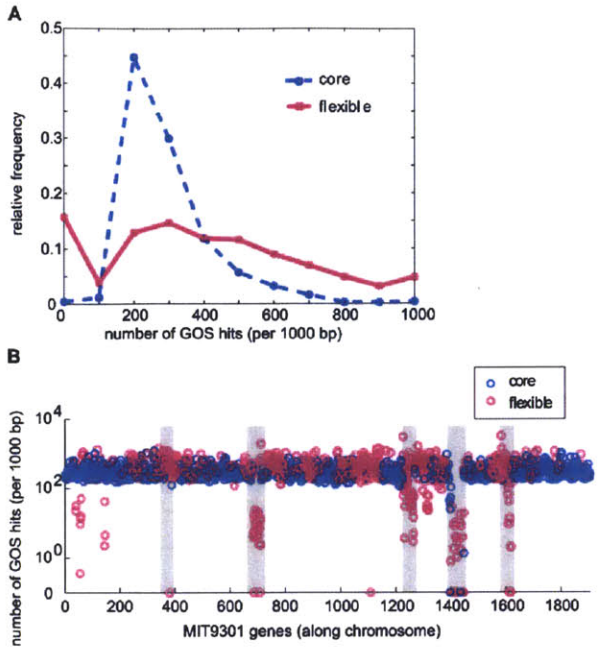


Figure 5. *Prochlorococcus* Core and Flexible Genes in the Global Ocean Survey (GOS) Dataset [64]

(A) Frequency distribution of GOS hits per gene, using genes in the *Prochlorococcus* MIT9301 genome as queries. Most core genes retrieve a similar number of GOS hits, as one would expect from single copy genes shared by all *Prochlorococcus*, resulting in a relatively tight frequency distribution. In contrast, flexible genes retrieve a broad range of GOS hits per gene, consistent with their scattered distribution among genomes.
(B) The number of GOS hits per gene, again using MIT9301 genes as queries, plotted against position along the chromosome. Shaded regions represent genomic islands, after [60]. Flexible genes with low representation in the GOS dataset tend to be located in genomic islands. In both (A) and (B), the number of GOS hits per gene is normalized to gene length and plotted as hits per gene, per 1,000 bp.
doi:10.1371/journal.pgen.0030231.g005

selection pressure via predators and phage and transporter composition, which likely plays a role in both defense from toxins and differences in nutrient availability. The latter is consistent with our earlier observation that genes involved in phosphorus acquisition are distributed among *Prochlorococcus* isolates not according to phylogeny, but rather the P concentrations in their ocean of origin [11]. However, despite the clear evidence for common gene gains and losses throughout the evolution of *Prochlorococcus*, we still observed a significant correlation between genome content and phylogeny. This suggests an important contribution of vertically inherited genes to the overall genome content that cannot be easily substituted through lateral gene transfer or lost altogether.

The core genome of *Prochlorococcus*, with 81% of the 1,273 genes having an inferred function, is now reasonably well understood and appears to encode a viable cell. That this could be circumscribed through the analysis of only 12 genomes is encouraging, and likely emerges from the reasonably small evolutionary distance between these isolates. The close agreement between manually curated core pathway reconstruction for one isolate [8], and the automatic reconstruction of the core metabolism shared by all 12 isolates in our study, promises to help streamline the analysis of new genomes. To date, discussions of minimal genomes to support life have focused on the set of genes that enable heterotrophic cells to replicate on rich organic media, where they benefit from nutrients that must have been synthesized by other organisms [65]. Here, however, we are approximating the minimum number of genes necessary to convert solar energy, carbon dioxide, and inorganic nutrients to living biomass.

The *Prochlorococcus* flexible genome is still only loosely defined, as over 70% of the orthologous groups in this category have no known homolog in MicrobesOnline and no inferred function. Moreover, as the last genomes are added to the analysis, they each add roughly 150 new genes to the *Prochlorococcus* pan-genome (Figure 1); thus it appears that the global pool of genes that are residing, at this moment, in a *Prochlorococcus* cell cannot even be approximated from this dataset. Therefore, one of the most daunting unanswered questions is: How many *Prochlorococcus* genotypes truly exist in the ocean, and what fraction of these has differential fitness at any point in time?

The level of diversity found in the flexible genes, and the steady increment of genes added to the *Prochlorococcus* pan genome with each new genome, suggests that we have barely begun to observe the extent of micro-diversity among *Prochlorococcus* in the ocean. Although the sequencing of 12 genomes represents one of the larger sequencing projects of closely related isolates to date, each isolate undoubtedly represents a subclade of a very large number of cells—especially considering the approximately 10^{25} *Prochlorococcus* cells in the ocean [3]. Additional sequencing, especially metagenomic [63] and single-cell sequencing [66], will help us understand more about on what scale, and where in the genomes, the flexible genes vary. In particular, it will be enlightening to understand the complete genome diversity of the 10^5 cells in a milliliter of ocean water, and conversely, how widely separated in space two cells with identical genomes might be.

Materials and Methods

DNA sequencing and assembly. The genome sequences of eight of the isolates used in our analysis are reported for the first time here. The genomes of MIT9211, MIT9515, NATL1A, MIT9303, MIT9301, and AS9601 were sequenced by the J. Craig Venter Institute as follows: Two genomic libraries with insert sizes of 4 and 40 kb were made as described in [67]. The prepared plasmid and fosmid clones were sequenced from both ends to provide paired-end reads at the J. Craig Venter Institute Joint Technology Center on ABI3730XL DNA sequencers (Applied Biosystems). Successful reads for each organism were used as input for the Celera Assembler. WGS sequence produced by the assembler was then annotated using the PGAAP at NCBI. Accession numbers for all genomes are provided in Table 1.

NATL2A was sequenced at the DOE Joint Genome Institute by methods described previously (http://www.jgi.doe.gov/sequencing/protocols/prots_production.html). Briefly, three whole genome shotgun libraries were constructed containing inserts of approximately 3 kb, 8 kb, or 40 kb and sequenced to a depth of 9X using BigDye Terminators on ABI3730 sequencersb (Applied Biosystems). Shotgun reads were assembled with parallel PHRAP (http://www.phrap.org).

The MIT9215 genome was sequenced with a combination of approximately 20X coverage of 454 pysoquencing (454 Life Sciences) and standard Sanger sequencing of 3-kb insert libraries. All genomes were completed to finished quality with no gaps, except MIT9211, with one gap of less than 1 kb and an estimated error rate of less than 1 in 50,000 bases.

Genome annotation. We re-annotated 12 sequenced *Prochlorococcus*

and four finished marine *Synechococcus* genomes by a uniform method for the purpose of this study. We used the gene prediction programs CRITICA [68] and GLIMMER [69]. The results from both programs were combined into a preliminary set of unique ORFs. Overlapping gene models from the two programs are considered the same gene if sharing the same stop position and in the same reading frame, in which case the gene start site of the CRITICA model is preferred. Coding genes that are shorter than 50 aa long are excluded unless they are conserved in more than one genome. Orthologous genes between two given genomes are assigned automatically using MicrobesOnline's [70] (http://www.microbesonline.org) genome annotation pipeline. The new annotations are also available at that site.

Two genes are considered orthologs if they are reciprocal best BLASTp hits and the alignment covers at least 75% of the length of each gene. An orthologous group includes all genes that are orthologous to any other gene in the group. The most common challenge of clustering orthologous genes is the risk of merging paralogous genes into one group. However, our method yields only 127 paralog-containing groups. In those cases, gene neighborhoods were also compared. Because a single missing ortholog effectively removes a gene from the core genome, the clusters that are absent in only one or two genomes were verified by BLAST search.

While the COG categories alone provide enough information to draw these conclusions about the membrane synthesis enzymes, there are some shortcomings. Some *Prochlorococcus* orthologous groups can be annotated with a gene name but not a COG (for example the LPS synthesis gene wcaK, or many photosystem genes like psbA), where literature searches show that they are likely involved in LPS synthesis. Other categories are hampered by the arrangement of the COG categories, which were not chosen with any particular focus on this system. For example, the category "Amino acid transport and metabolism" includes transporters and intracellular enzymes. When we found that transporters are among the most recently gained genes, we desired a way to group all of them by themselves. We decided the best approach was to group genes into five broad categories on the basis of keyword searches: membrane or cell wall synthesis, transporters, photosynthesis, DNA repair or modification, and other. HLI proteins were identified by their possession of six out of ten conserved residues in the motif AExxNGRxAMIGF, and lengths under 120 amino acids [32].

Phylogenetic analysis. 16S rRNA and 16S-23S rRNA ITS region sequences were manually aligned in ARB and phylogenetic reconstruction using maximum parsimony, neighbor-joining, and maximum likelihood was done in PAUP [71]. Following the approach described in [33] to identify the phylogenetic relationship between the sequenced isolates, we aligned all core genes using clustalw using the protein sequence as reference. We randomly concatenated 100 alignments and constructed a phylogenetic tree using maximum parsimony and bootstrap resampled 100 times. The random concatenation was repeated 100 times and the average bootstrap values for concatenated alignments are reported in Figure 2. In addition, we also constructed a phylogenetic tree using maximum parsimony on each individual alignment and the most likely tree for each gene (plurality consensus tree based on 100 bootstraps) was identified. We also calculated the phylogenetic relationship based on the presence and absence of orthologous groups as previously described [34]. However, we used bootstrap instead of jack-knife resampling to test how well individual nodes were supported to ensure easy comparison with other phylogenetic trees.

Estimation of the timing of gene loss and gain events was as described using a maximum parsimony approach [35]. We used the phylogenetic tree in Figure 2C rooted between the *Prochlorococcus* and *Synechococcus* last common ancestors as a guide. We included the cost of a "gain" event in the tree's common ancestor node. We assigned a gene gain event twice the cost of a loss event, and in cases where two scenarios had equal scores we chose the one with fewer gains. We also tested a ratio of three to one, which changes the behavior of 117 genes.

Metabolic reconstruction. To predict the metabolic pathways present in the sequenced isolates, we ran Pathway Tools software [30] to generate a Pathway/Genome database (PGDB). This software creates gene, protein, reaction, small-molecule, and pathway objects based on Enzyme Commission (E.C.) numbers and enzyme names assigned in the genome annotation. We hand-curated the PGDB to eliminate unlikely pathways, and from it we created a pathway model of the central carbon metabolism [72]. To aid in the analysis of the core and flexible genes, we created a pseudogenome, Pan, which includes all genes from all isolates. We created another pseudogenome for the core genome. The database is available in flat file, BioPAX, and SBML format.

Supporting Information

Figure S1. The Core Genome Includes Enzymes for Central Carbon Metabolism, Including the Calvin Cycle, Glycolysis, and an Incomplete TCA Cycle Producing Fumarate and 2-Oxoglutarate

Some genomes, but not the core genome, also include *sdhAB*, encoding an enzyme for the reaction 1.3.99.1, the conversion of fumarate to succinate (Table 2). The pathway diagram includes the stuctures of intermediate metabolites, the locus name, in MED4, of the gene encoding each enzyme, the enzyme name, and the E.C. number.

Found at doi:10.1371/journal.pgen.0030231.sg001 (1.4 MB EPS).

Figure S2. The Core Genome Includes Enzymes for the Synthesis of All 20 Amino Acids

The pathway diagram is annotated as in Figure S1.

Found at doi:10.1371/journal.pgen.0030231.sg002 (2.2 MB EPS).

Figure S3. The Core Genome Includes Enzymes for the Synthesis of Divinyl Chlorophyll

The pathway diagram is annotated as in Figure S1.

Found at doi:10.1371/journal.pgen.0030231.sg003 (1.5 MB EPS).

Figure S4. The Core Genome Includes Enzymes for the Synthesis of the Cofactors NAD (A), Coenzyme A (B and C), and FAD (D)

The pathway diagrams are annotated as in Figure S1. One reaction (2.7.1.33) in coenzyme A synthesis is highlighted; its enzyme (pantothenate kinase) has not been identified in the core or pan-genomes.

Found at doi:10.1371/journal.pgen.0030231.sg004 (1.3 MB EPS).

Figure S5. Islands of LL Genomes Not Represented in Figure 4

The dot plot shows the location of each gene, the ancestor in which it is estimated to be acquired, and when possible, the best match outside *Prochlorococcus*. The lower plot is the number of genes gained in a sliding window (size 10,000 bp, interval 1,000 bp) along the chromosome.

Found at doi:10.1371/journal.pgen.0030231.sg005 (4.2 MB EPS).

Figure S6. Islands of HL Genomes Not Represented in Figure 4

The dot plot shows the location of each gene, the ancestor in which it is estimated to be acquired, and when possible, the best match outside *Prochlorococcus*. The lower plot is the number of genes gained in a sliding window (size 10,000 bp, interval 1,000 bp) along the chromosome. When available, the locations of islands previously defined by hand are represented by shaded regions.

Found at doi:10.1371/journal.pgen.0030231.sg006 (6.1 MB EPS).

Table S1. All *Prochlorococcus* Orthologous Groups in This Study

For each group, its locus names are given for those genomes in which it is found. Also given are the COG match [55], gene name, and description as assigned by MicrobesOnline (http://www.microbesonline.org).

Found at doi:10.1371/journal.pgen.0030231.st001 (1.8 MB XLS).

Table S2. Prochlorococcus Core Genes Absent in Synechococcus

33 orthologous groups are shared by all *Prochlorococcus* but absent in some *Synechococcus*, and only 13 of those are absent in all *Synechococcus*. For each such orthologous group, its presence or absence in each of the four *Synechococcus* genomes in this analysis is given. Also given is the locus name for the gene in MED4, its COG match, and its gene name, if available.

Found at doi:10.1371/journal.pgen.0030231.st002 (68 KB DOC).

Table S3. Genes Found in All Synechococcus but No Prochlorococcus

The locus name for *Synechococcus* is given, in addition to the COG and gene name, if available.

Found at doi:10.1371/journal.pgen.0030231.st003 (45 KB XLS).

Table S4. Genes Lost or Gained at Each Ancestor

For each gene, the name and COG are given, in addition to a locus name. The role assigned is one of "nomatch," "shortnomatch," "conserved_unknown," "hli," "photosynthesis," "DNA," "membrane," "transport," or "other," on the basis of keyword matches in the gene name, COG, or description. The latter five categories are

reported individually in Figure 3B; the totals are reported in Figure 3A.

Found at doi:10.1371/journal.pgen.0030231.st004 (1.5 MB XLS).

Table S5. The Most Common COGs in the Core and Flexible Genomes

We used matches against the COG database as a first impression of the differences between the core and flexible genomes. The number of *Prochlorococcus* orthologous groups and the total number of genes in those groups, matching each COG is given. The top ten COGs matching the core and flexible genomes are shown.

Found at doi:10.1371/journal.pgen.0030231.st005 (43 KB DOC).

Table S6. Orthologous Groups Found in All HL Isolates

These include those exclusive to HL isolates and those shared with some, but not all, LL isolates, as indicated. Also given are the gene name, description, and COG assignments as in Table S1.

Found at doi:10.1371/journal.pgen.0030231.st006 (114 KB XLS).

Table S7. Orthologous Groups Found in All LL Isolates

As Table S6, but those found in all LL isolates.

Found at doi:10.1371/journal.pgen.0030231.st007 (60 KB XLS).

Table S8. Notable Genes Exclusive to eMIT9313 Isolates

These are orthologous groups from Table S1, each found only in MIT9303, MIT9313, and in some cases marine *Synechococcus*. This list includes only those genes with hypothetical functions and with no BLAST alignment against the other genomes. Note that some belong to COGs shared with other *Prochlorococcus* isolates, but their extreme sequence divergence suggests their precise roles differ.

Found at doi:10.1371/journal.pgen.0030231.st008 (118 KB XLS).

Acknowledgments

Sequencing, assembly, and annotation efforts of MIT9211, MIT9515, NATL1A, MIT9303, MIT9301, and AS9601 were supported by Marine Microbial Sequencing Project of the Gordon and Betty Moore Foundation, and carried out at the J. Craig Venter Institute (JCVI) Joint Technology Center, under the leadership of Robert Friedman and Yu-Hui Rogers. We thank Granger Sutton, Aaron Halpern, and Saul A. Kravitz for their contributions to completion of these genomes. Sequencing, assembly, and annotation of the genomes NATL2A and MIT9215 were produced by the DOE Joint Genome Institute JGI. We thank Brian Palenik for allowing us to use unpublished *Synechococcus* genomes as out-group reference genomes in our analyses.

Author contributions. ACM, GMC, and SWC conceived and designed the experiments. ACM, SR, FC, AL, SF, JJ, CS, and PR performed the experiments. GCK, ACM, KH, JZ, and MLC analyzed the data. GCK, ACM, MLC, and SWC wrote the paper.

Funding. This work was supported in part by grants from the National Science Foundation (NSF), Department of Energy (DOE), and the Gordon and Betty Moore Foundation (SWC), and a DOE-GTL Grant (SWC and GC). It is C-MORE Contribution #43. GCK was supported in part by a National Institutes of Health Training Grant through the MIT Biology Department, MLC by an NSF Graduate Fellowship, and ACM by a fellowship from the Danish National Science Foundation.

Competing interests. The authors have declared that no competing interests exist.

References

1. Goericke RE, Welschmeyer NA (1993) The marine prochlorophyte *Prochlorococcus* contributes significantly to phytoplankton biomass and primary production in the Sargasso Sea. Deep Sea Research (Part I, Oceanographic Research Papers) 40: 2283–2294.
2. Waterbury JB, Watson SW, Valois FW, Franks DG (1986) Biological and ecological characterization of the marine unicellular bacterium *Synechococcus*. Can Bull Fish Aquat Sci 214: 71–120.
3. Partensky F, Hess WR, Vaulot D (1999) *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. Microbiol Mol Biol Rev 63: 106–127.
4. Moore LR, Rocap G, Chisholm SW (1998) Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. Nature 393: 464–467.
5. West NJ, Scanlan DJ (1999) Niche-partitioning of *Prochlorococcus* populations in a stratified water column in the eastern North Atlantic Ocean. Appl Environ Microbiol 65: 2585–2591.
6. Urbach E, Scanlan DJ, Distel DL, Waterbury JB, Chisholm SW (1998) Rapid diversification of marine picophytoplankton with dissimilar light-harvesting structures inferred from sequences of *Prochlorococcus* and *Synechococcus* (Cyanobacteria). J Mol Evol 46: 188–201.
7. Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, et al. (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature 424: 1042–1047.
8. Dufresne A, Salanoubat M, Partensky F, Artiguenave F, Axmann IM, et al. (2003) Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. Proc Natl Acad Sci U S A 100: 10020–10025.
9. Palenik B, Ren Q, Dupont CL, Myers GS, Heidelberg JF, et al. (2006) Genome sequence of *Synechococcus* CC9311: insights into adaptation to a coastal environment. Proc Natl Acad Sci U S A 103: 13555–13559.
10. Moore LR, Goericke RE, Chisholm SW (2002) Utilization of different nitrogen sources by the marine cyanobacteria, *Prochlorococcus* and *Synechococcus*. Limnol Oceanogr 47: 989–996.
11. Martiny AC, Coleman ML, Chisholm SW (2006) Phosphate acquisition genes in *Prochlorococcus* ecotypes: evidence for genome-wide adaptation. Proc Natl Acad Sci U S A 103: 12552–12557.
12. Bapteste E, Boucher Y, Leigh J, Doolittle WF (2004) Phylogenetic reconstruction and lateral gene transfer. Trends Microbiol 12: 406–411.
13. Lerat E, Daubin V, Moran NA (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-proteobacteria. PLoS Biol 1: e19. doi:10.1371/journal.pbio.0000019
14. Doolittle WF (1999) Phylogenetic classification and the universal tree. Science 284: 2124–2129.
15. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, et al. (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome." Proc Natl Acad Sci U S A 102: 13950–13955.
16. Makarova K, Slesarev A, Wolf Y, Sorokin A, Mirkin B, et al. (2006) Comparative genomics of the lactic acid bacteria. Proc Natl Acad Sci U S A 103: 15611–15616.
17. Mulkidjanian AY, Koonin EV, Makarova KS, Mekhedov SL, Sorokin A, et al. (2006) The cyanobacterial genome core and the origin of photosynthesis. Proc Natl Acad Sci U S A 103: 13126–13131.
18. Rocap G, Distel DL, Waterbury JB, Chisholm SW (2002) Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. Appl Environ Microbiol 68: 1180–1191.
19. Moore LR, Chisholm SW (1999) Photophysiology of the marine Cyanobacterium *Prochlorococcus*: ecotypic differences among cultured isolates. Limnol Oceanogr 44: 628–638.
20. Partensky F, Hoepffner N, Li W, Ulloa O, Vaulot D (1993) Photoacclimation of *Prochlorococcus* sp. (Prochlorophyta) strains Isolated from the North Atlantic and the Mediterranean Sea. Plant Physiol 101: 285–296.
21. Shalapyonok A, Olson RJ, Shalapyonok LS (1998) Ultradian Growth in *Prochlorococcus* spp. Appl Environ Microbiol 64: 1066–1069.
22. Scanlan DJ, Hess WR, Partensky F, Vaulot D (1996) High degree of genetic variation in *Prochlorococcus* (Prochlorophyta) revealed by RFLP analysis. European Journal of Phycology 31: 1–9.
23. Lawrence JG, Hendrickson H (2005) Genome evolution in bacteria: order beneath chaos. Curr Opin Microbiol 8: 572–578.
24. Steglich C, Futschik M, Rector T, Steen R, Chisholm SW (2006) Genome-wide analysis of light sensing in *Prochlorococcus*. J Bacteriol 188: 7796–7806.
25. Nagata N, Tanaka R, Satoh S, Tanaka A (2005) Identification of a vinyl reductase gene for chlorophyll synthesis in *Arabidopsis thaliana* and implications for the evolution of Prochlorococcus species. Plant Cell 17: 233–240.
26. Chisholm SW, Frankel SL, Goericke RE, Olson RJ, Palenik B, et al. (1992) *Prochlorococcus marinus* nov. gen. nov. sp.: an oxyphototrophic marine prokaryote containing divinyl chlorophyll *a* and *b*. Archives of Microbiology 157: 297–300.
27. Rubio LM, Flores E, Herrero A (1999) Molybdopterin guanine dinucleotide cofactor in *Synechococcus* sp. nitrate reductase: identification of *mobA* and isolation of a putative *moeB* gene. FEBS Lett 462: 358–362.
28. Rubio LM, Flores E, Herrero A (2002) Purification, cofactor analysis, and site-directed mutagenesis of *Synechococcus* ferredoxin-nitrate reductase. Photosynth Res 72: 13–26.
29. Lomas MW, Lipschultz F (2006) Forming the primary nitrite maximum: nitrifiers or phytoplankton? Limnol Oceanogr 51: 2453–2467.
30. Paley SM, Karp PD (2006) The pathway tools cellular overview diagram and omics viewer. Nucleic Acids Res 34: 3771–3778.
31. Ferris MJ, Palenik B (1998) Niche adaptation in ocean cyanobacteria. Nature 396: 226–228.
32. Lindell D, Sullivan MB, Johnson ZI, Tolonen AC, Rohwer F, et al. (2004) Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. Proc Natl Acad Sci U S A 101: 11013–11018.
33. Rokas A, Williams BL, King N, Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425: 798–804.

34. Snel B, Bork P, Huynen MA (1999) Genome phylogeny based on gene content. Nat Genet 21: 108–110.
35. Mirkin BG, Fenner TI, Galperin MY, Koonin EV (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor, and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol Biol 3: 2.
36. Palenik B, Brahamsha B, Larimer FW, Land M, Hauser L, et al. (2003) The genome of a motile marine *Synechococcus*. Nature 424: 1037–1042.
37. Ahlgren NA, Rocap G, Chisholm SW (2006) Measurement of *Prochlorococcus* ecotypes using real-time polymerase chain reaction reveals different abundances of genotypes with similar light physiologies. Environ Microbiol 8: 441–454.
38. Moore LR, R. G, S.W. C (1995) Comparative physiology of *Synechococcus* and *Prochlorococcus*: influence of light and temperature on growth, pigments, fluorescence and absorptive properties. Marine Ecology Progress Series 116: 259–275.
39. Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EM, et al. (2006) Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. Science 311: 1737–1740.
40. Zinser ER, Johnson ZI, Coe A, Karaca E, Veneziano D, et al. (2007) Influence of light and temperature on *Prochlorococcus* ecotype distribution in the Atlantic Ocean. Limnol Oceanogr 52: 2205–2220.
41. Bouman HA, Ulloa O, Scanlan DJ, Zwirglmaier K, Li WK, et al. (2006) Oceanographic basis of the global surface distribution of *Prochlorococcus* ecotypes. Science 312: 918–921.
42. West NJ, Schonhuber WA, Fuller NJ, Amann RI, Rippka R, et al. (2001) Closely related *Prochlorococcus* genotypes show remarkably different depth distributions in two oceanic regions as revealed by in situ hybridization using 16S rRNA-targeted oligonucleotides. Microbiology 147: 1731–1744.
43. Tolonen AC, Aach J, Lindell D, Johnson ZI, Rector T, et al. (2006) Global gene expression of *Prochlorococcus* ecotypes in response to changes in nitrogen availability. Mol Syst Biol 2: 53.
44. Lindell D, Jaffe JD, Coleman ML, Futschik ME, Axmann IM, et al. (2007) Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. Nature 449: 83–86.
45. Sasaki Y, Ishikawa J, Yamashita A, Oshima K, Kenri T, et al. (2002) The complete genomic sequence of *Mycoplasma penetrans*, an intracellular bacterial pathogen in humans. Nucleic Acids Res 30: 5293–5300.
46. Park JH, Burns K, Kinsland C, Begley TP (2004) Characterization of two kinases involved in thiamine pyrophosphate and pyridoxal phosphate biosynthesis in Bacillus subtilis: 4-amino-5-hydroxymethyl-2methylpyrimidine kinase and pyridoxal kinase. J Bacteriol 186: 1571–1575.
47. Toms AV, Haas AL, Park JH, Begley TP, Ealick SE (2005) Structural characterization of the regulatory proteins TenA and TenI from *Bacillus subtilis* and identification of TenA as a thiaminase II. Biochemistry 44: 2319–2329.
48. Hess WR, Rocap G, Ting CS, Larimer F, Stilwagen S, et al. (2001) The photosynthetic apparatus of *Prochlorococcus*: insights through comparative genomics. Photosynth Res 70: 53–71.
49. Steglich C, Mullineaux CW, Teuchner K, Hess WR, Lokstein H (2003) Photophysical properties of *Prochlorococcus marinus* SS120 divinyl chlorophylls and phycoerythrin in vitro and in vivo. FEBS Lett 553: 79–84.
50. Steglich C, Frankenberg-Dinkel N, Penno S, Hess WR (2005) A green light-absorbing phycoerythrin is present in the high-light-adapted marine cyanobacterium *Prochlorococcus* sp. MED4. Environ Microbiol 7: 1611–1618.
51. Zubkov MV, Fuchs BM, Tarran GA, Burkill PH, Amann R (2003) High rate of uptake of organic nitrogen compounds by *Prochlorococcus* cyanobacteria as a key to their dominance in oligotrophic oceanic waters. Appl Environ Microbiol 69: 1299–1304.
52. Lu AL, Li X, Gu Y, Wright PM, Chang DY (2001) Repair of oxidative DNA damage: mechanisms and functions. Cell Biochem Biophys 35: 141–170.
53. Dufresne A, Garczarek L, Partensky F (2005) Accelerated evolution associated with genome reduction in a free-living prokaryote. Genome Biol 6: R14.
54. Chavez S, Lucena JM, Reyes JC, Florencio FJ, Candau P (1999) The presence of glutamate dehydrogenase is a selective advantage for the Cyanobacterium *Synechocystis* sp. strain PCC 6803 under nonexponential growth conditions. J Bacteriol 181: 808–813.
55. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, et al. (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4: 41.
56. Zinser ER, Coe A, Johnson ZI, Martiny AC, Fuller NJ, et al. (2006) *Prochlorococcus* ecotype abundances in the North Atlantic Ocean as revealed by an improved quantitative PCR method. Appl Environ Microbiol 72: 723–732.
57. Pal C, Papp B, Lercher MJ (2005) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet 37: 1372–1375.
58. Bibby TS, Mary I, Nield J, Partensky F, Barber J (2003) Low-light-adapted *Prochlorococcus* species possess specific antennae for each photosystem. Nature 424: 1051–1054.
59. Sullivan MB, Lindell D, Lee JA, Thompson LR, Bielawski JP, et al. (2006) Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. PLoS Biol 4: e234. doi:10.1371/journal.pbio.0040234
60. Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, et al. (2006) Genomic islands and the ecology and evolution of *Prochlorococcus*. Science 311: 1768–1770.
61. Adir N, Zer H, Shochat S, Ohad I (2003) Photoinhibition: a historical perspective. Photosynth Res 76: 343–370.
62. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304: 66–74.
63. DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, et al. (2006) Community genomics among stratified microbial assemblages in the ocean's interior. Science 311: 496–503.
64. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, et al. (2007) The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. PLoS Biol 5: e77. doi:10.1371/journal.pbio.0050016
65. Gil R, Silva FJ, Pereto J, Moya A (2004) Determination of the core of a minimal bacterial gene set. Microbiol Mol Biol Rev 68: 518–537.
66. Zhang K, Martiny AC, Reppas NB, Barry KW, Malek J, et al. (2006) Sequencing genomes from single cells by polymerase cloning. Nat Biotechnol 24: 680–686.
67. Goldberg SM, Johnson J, Busam D, Feldblyum T, Ferriera S, et al. (2006) A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. Proc Natl Acad Sci U S A 103: 11240–11245.
68. Badger JH, Olsen GJ (1999) CRITICA: coding region identification tool invoking comparative analysis. Mol Biol Evol 16: 512–524.
69. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL (1999) Improved microbial gene identification with GLIMMER. Nucleic Acids Res 27: 4636–4641.
70. Alm EJ, Huang KH, Price MN, Koche RP, Keller K, et al. (2005) The MicrobesOnline Web site for comparative genomics. Genome Res 15: 1015–1022.
71. Swofford DL (2003) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. 4 ed. Sunderland, Massachusetts: Sinauer Associates.
72. Segre D, Zucker J, Katz J, Lin X, D'Haeseleer P, et al. (2003) From annotated genomes to metabolic flux models and kinetic parameter fitting. Omics 7: 301–316.

Supporting information for chapter 2

Some supplemental figures and tables are too large to print here but are available at the paper's PLOS Genetics website:

`http://www.plosgenetics.org/article/info:doi/10.1371/journal.pgen.0030231`

Figures S1-S4 Available at PLOS Genetics website

Figures S5 and S6 Below

Table S1 Available at PLOS Genetics website

Table S2 Below

Tables S3-S4 Available at PLOS Genetics website

Table S5 Below

Tables S6-S8 Available at PLOS Genetics website

SynWH8102	SynCC9605	SynCC9902	SynCC9311	MED4 locus	COG	gene name
0	0	0	0	PMED4_00681		
0	0	0	0	PMED4_02181		
0	0	0	0	PMED4_06781	COG786:Na+/glutamate symporter [Amino acid transport and metabolism]	GltS
0	0	0	0	PMED4_06861	COG1535:Isochorismate hydrolase [Secondary metabolites biosynthesis, transport, and catabolism]	EntB/pncA
0	0	0	0	PMED4_07061		
0	0	0	0	PMED4_08191		
0	0	0	0	PMED4_08941	COG2146:Ferredoxin subunits of nitrite reductase and ring-hydroxylating dioxygenases [Inorganic ion transport and metabolism / General function prediction only]	NirD / hcaE
0	0	0	0	PMED4_11001		
0	0	0	0	PMED4_11581		
0	0	0	0	PMED4_11671		
0	0	0	0	PMED4_12731		
0	0	0	0	PMED4_15181		
0	0	0	0	PMED4_15741		hli11

Table 2.1: Table S2: *Prochlorococcus* Core Genes Absent in *Synechococcus*. 33 orthologous groups are shared by all *Prochlorococcus* but absent in some Synechococcus, and only 13 of those are absent in all Synechococcus. For each such orthologous group, its presence or absence in each of the four Synechococcus genomes in this analysis is given. Also given is the locus name for the gene in MED4, its COG match, and its gene name, if available.

SynWH8102	SynCC9605	SynCC9902	SynCC9311	MED4 locus	COG	gene name
1	0	0	1	PMED4_00811	COG2091:Phosphopantetheinyl transferase [Coenzyme metabolism]	Sfp
0	1	1	1	PMED4_02171	COG492:Thioredoxin reductase [Posttranslational modification, protein turnover, chaperones]	TrxB
0	0	0	1	PMED4_02191	COG3329:Predicted permease [General function prediction only]	sbtA
1	1	0	0	PMED4_02251		
1	1	0	0	PMED4_03481		
0	0	0	1	PMED4_03761	COG5470:Uncharacterized conserved protein [Function unknown]	
1	1	1	0	PMED4_05531	COG1324:Uncharacterized protein involved in tolerance to divalent cations [Inorganic ion transport and metabolism]	cutA
0	1	1	1	PMED4_06771		pcbA
1	0	0	0	PMED4_07161		
0	0	1	0	PMED4_07701		
0	1	1	1	PMED4_08901	COG1528:Ferritin-like protein [Inorganic ion transport and metabolism]	/ Ftn
0	1	1	1	PMED4_08921	COG664:cAMP-binding proteins - catabolite gene activator and regulatory subunit of cAMP-dependent protein kinases [Signal transduction mechanisms]	Crp
1	1	1	0	PMED4_11181		
0	1	1	1	PMED4_11231		
0	0	0	1	PMED4_12131		
1	1	0	1	PMED4_12251		
0	1	1	1	PMED4_13361	COG716:Flavodoxins [Energy production and conversion]	FldA
0	1	0	1	PMED4_14561	COG63:Predicted sugar kinase [Carbohydrate transport and metabolism] / COG62:Uncharacterized conserved protein [Function unknown]	
1	1	1	0	PMED4_15631		
0	0	0	1	PMED4_18811		

Table 2.2: Table S2, continued.

COG Name	COG Description	*Prochlorococcus* Orthologous Groups	genes
Core			
COG524	Sugar kinases, ribokinase family [Carbohydrate transport and metabolism]	3	42
COG456	Acetyltransferases [General function prediction only]	3	42
COG1132	ABC-type multidrug transport system, ATPase and permease components [Defense mechanisms]	4	56
COG740	Protease subunit of ATP-dependent Clp proteases [Posttranslational modification, protein turnover, chaperones / Intracellular trafficking and secretion]	4	56
COG745	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain [Signal transduction mechanisms / Transcription]	4	56
COG596	Predicted hydrolases or acyltransferases (alpha/beta hydrolase superfamily) [General function prediction only]	4	56
COG1028	Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases) [Secondary metabolites biosynthesis, transport, and catabolism / General function prediction only]	4	56
COG465	ATP-dependent Zn proteases [Posttranslational modification, protein turnover, chaperones]	4	56
COG568	DNA-directed RNA polymerase, sigma subunit (sigma70/sigma32) [Transcription]	5	70
COG451	Nucleoside-diphosphate-sugar epimerases [Cell envelope biogenesis, outer membrane / Carbohydrate transport and metabolism]	6	112

Table 2.3: Table S5: The Most Common COGs in the Core and Flexible Genomes. We used matches against the COG database as a first impression of the differences between the core and flexible genomes. The number of *Prochlorococcus* orthologous groups and the total number of genes in those groups, matching each COG is given. The top ten COGs matching the core and flexible genomes are shown.

COG Name	COG Description	*Prochlorococcus* Orthologous Groups	genes
COG457	FOG: TPR repeat [General function prediction only]	7	8
COG1028	Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases) [Secondary metabolites biosynthesis, transport, and catabolism / General function prediction only]	7	20
COG2274	ABC-type bacteriocin/lantibiotic exporters, contain an N-terminal double-glycine peptidase domain [Defense mechanisms]	7	20
COG463	Glycosyltransferases involved in cell wall biogenesis [Cell envelope biogenesis, outer membrane]	8	25
COG1943	Transposase and inactivated derivatives [DNA replication, recombination, and repair]	10	11
COG399	Predicted pyridoxal phosphate-dependent enzyme apparently involved in regulation of cell wall biogenesis [Cell envelope biogenesis, outer membrane]	11	16
COG5010	Flp pilus assembly protein TadD, contains TPR repeats [Intracellular trafficking and secretion]	12	20
COG451	Nucleoside-diphosphate-sugar epimerases [Cell envelope biogenesis, outer membrane / Carbohydrate transport and metabolism]	13	33
COG3063	Tfp pilus assembly protein PilF [Cell motility and secretion / Intracellular trafficking and secretion]	17	24
COG438	Glycosyltransferase [Cell envelope biogenesis, outer membrane]	25	89

Table 2.4: Table S5, continued.

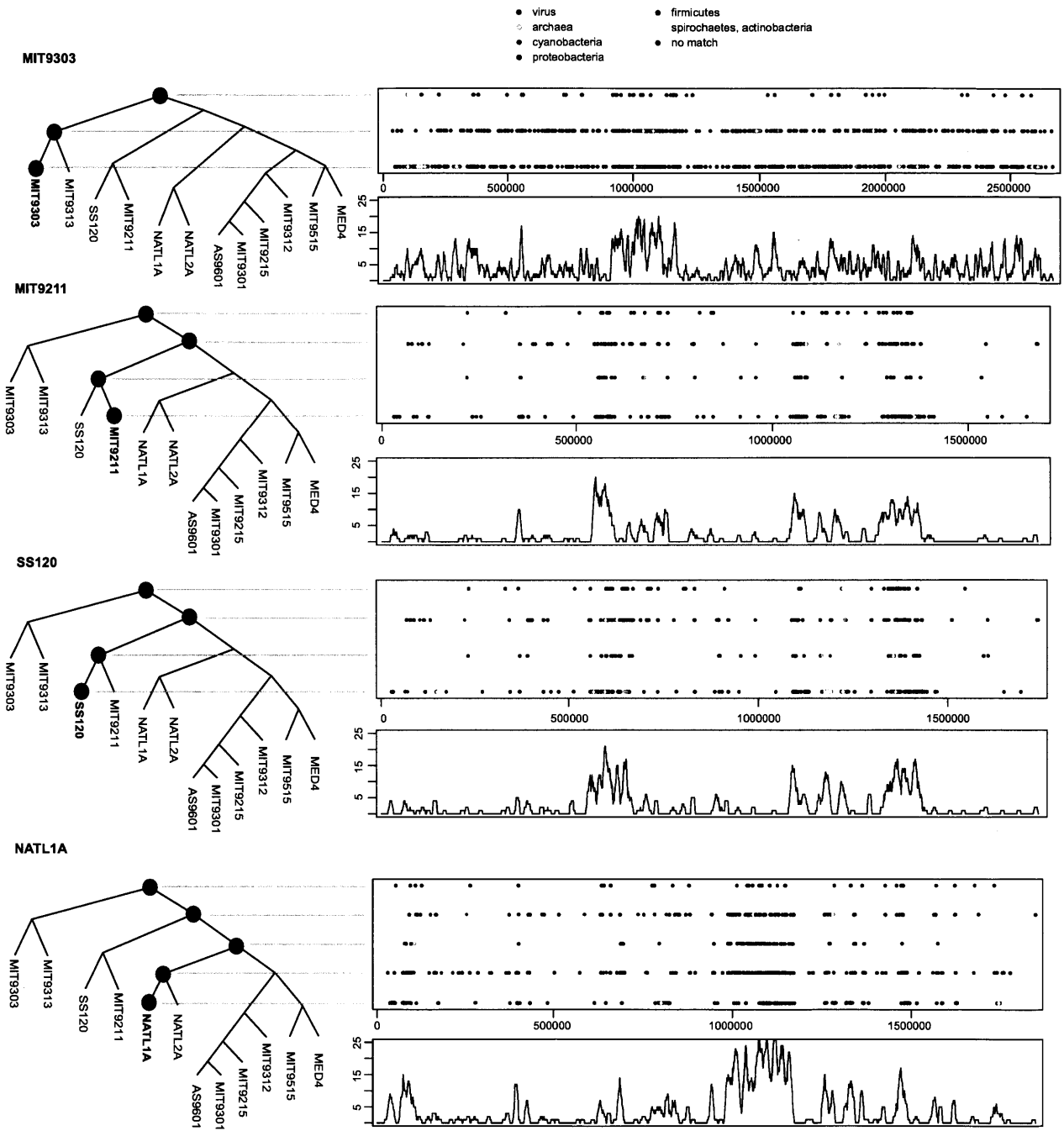


Figure 2-1: Figure S5: Islands of LL Genomes Not Represented in Figure 4. The dot plot shows the location of each gene, the ancestor in which it is estimated to be acquired, and when possible, the best match outside *Prochlorococcus*. The lower plot is the number of genes gained in a sliding window (size 10,000 bp, interval 1,000 bp) along the chromosome.

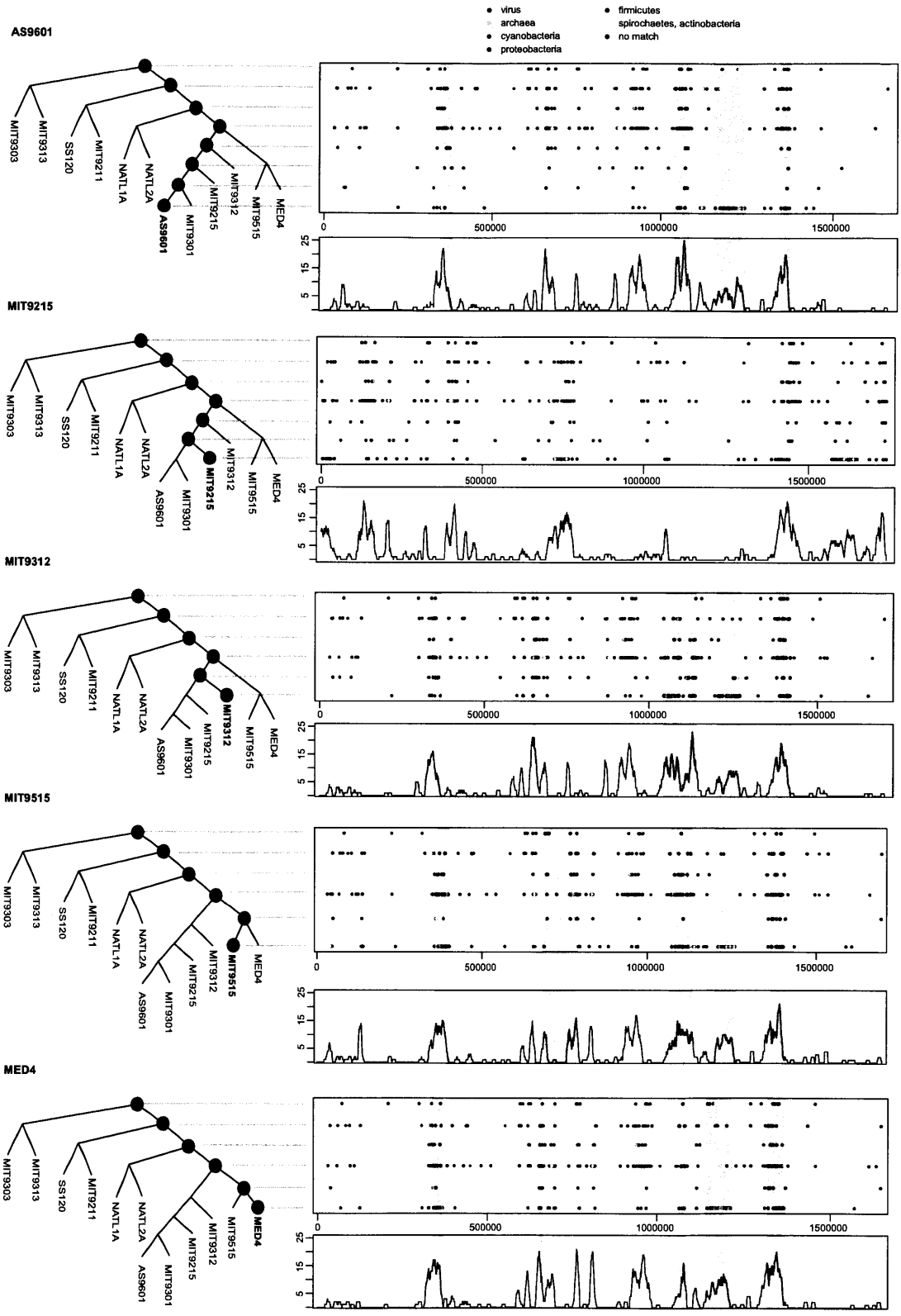


Figure 2-2: Figure S6: Islands of HL Genomes Not Represented in Figure 4. The dot plot shows the location of each gene, the ancestor in which it is estimated to be acquired, and when possible, the best match outside *Prochlorococcus*. The lower plot is the number of genes gained in a sliding window (size 10,000 bp, interval 1,000 bp) along the chromosome. When available, the locations of islands previously defined by hand are represented by shaded regions.

Chapter 3

Differences in timing and magnitude of *Prochlorococcus* ecotypes' responses to abrupt increases in light intensity

Abstract

Prochlorococcus thrives throughout the top 200m of the open ocean water column, in part because different *Prochlorococcus* ecotypes are adapted to growth at different light intensities. Much of that adaptation took place early in *Prochlorococcus* evolution as the high light-adapted (HL) ecotype diverged from the low light-adapted (LL). However, there is variation in light adaptation within the LL ecotype as well. The LL NATL clade (eNATL) has previously been detected in large numbers near the surface during deep mixing events, unlike other LL clades. This different behavior is thought to be due to eNATL's greater ability to survive short exposures to sudden increases in light intensity. The mechanism of this tolerance is not known, but the eNATL genomes' exceptional number of *hli* genes encoding high light inducible proteins (HLIPs) may play a role. Here, the early response and recovery of *Prochlorococcus* isolates is examined through fluorescence-based and flow cytometric methods. While all HL and LL isolates appear to be stressed by light shocks at various intensities, the HL and eNATL isolates respond more quickly in quenching their chlorophyll fluorescence and return to normal growth within 24 hours of a moderate high light shock. Other LL isolates exhibit a slower initial response but do not return to normal growth except after the mildest high light shocks. These results point to this early quenching as an explanation for eNATL's and HL isolates' robustness under such shocks. In the case of eNATL, this early quenching would be consistent with some proposed mechanisms of HLIPs'

interaction with the photosystem.

3.1 Introduction

Prochlorococcus is the most abundant cyanobacterium on Earth because it colonizes a very large habitat: the upper 200m of the open ocean. While this habitat has a number of consistent properties, enough variation exists within it that it might better be considered as representing a spectrum of niche dimensions which overlap to create an endless variety of possible niches. For example, while low nutrient concentrations are a common challenge for any microbe in the open ocean, the identity of the limiting nutrient (that in least supply relative to the cell's requirements) can vary. For example, iron limits growth in parts of the Pacific where influx of iron dust is low relative to in the Atlantic (Duce and Tindale, 1991; Thompson et al., 2011; Rusch et al., 2010). With iron more available in the Atlantic, phosphorus becomes the limiting nutrient (Wu et al., 2000; Martiny et al., 2006, 2009a; Coleman and Chisholm, 2010).

Prochlorococcus, as a photoautotroph, faces other challenges. Because it grows throughout the upper 200 meters of the water column, it must also cope with a wide range of light intensities (Zinser et al., 2007; Johnson et al., 2006) (Malmstrom et al., 2010, Appendix C). Much of that range is addressed by genetic adaptation, as optimum light intensities for growth are the defining feature of the two greatest *Prochlorococcus* clades (Moore and Chisholm, 1999). Not surprisingly, the high light-adapted ecotypes are most abundant in the mixed layer though they also occupy the stratified deeper ocean, while the low light-adapted are rarely found in the mixed layer, existing instead in the deeper ocean (Johnson et al., 2006).

But while each cell has an optimal light intensity for growth (Moore and Chisholm, 1999), it will experience a range of intensities throughout its life. Light intensity changes dramatically from night to day, but the predictability of the day/night cycle means *Prochlorococcus* can synchronize itself to that cycle with a circadian clock (the *kaiBC* mechanism, a simpler form of the *kaiABC* cycle found in other cyanobacteria) (Axmann et al., 2009). This allows *Prochlorococcus* to initiate the transcription of certain genes slightly before their products are needed (Zinser et al., 2009). Light also varies unpredictably, from the cell's point of view, during mixing events. These are seasonal events during which the ocean mixed layer, usually limited to the top 30-50 meters of the water column, extends down to 200 meters or more. Mixing varies across different ocean sites, for example being more pronounced at the Bermuda Atlantic Time Series station (BATS) than

at the Hawaii Ocean Timeseries (HOT) (Steinberg et al., 2001; Karl and Lukas, 1996; Bingham and Lukas, 1996). Due to mixing, a cell must respond to rapid changes in light intensity. A *Prochlorococcus* cell growing at the lower extreme of the mixed layer may be carried to the surface in hours (Delhez and Deleersnijder, 2010; Lande and Wood, 1987; Denman and Gargett, 1983). Therefore, the ability to survive temporary, large increases in light intensity becomes as important as any genetic predisposition for growth at a particular intensity.

While photosynthetic organisms depend on sunlight for their existence, an excess has a paradoxical effect: photoinhibition, the slowing of photosynthesis as light intensity increases past a certain point (Falkowski and Raven, 2007). This is due to damage to the photosystem, its proteins or chlorophyll, by excess reactive oxygen species generated by the photosystem. Particularly vulnerable is the core photosystem protein PsbA (or D1), which even in a healthy cell degrades with a half-life of about 30 minutes (Long et al., 1994; Adir et al., 2003) . Photoinhibition occurs when the rate of damage to PsbA exceeds the rate of replacement. This can be due either to an increase in the rate of damage, or the inhibition of the synthesis of new proteins including PsbA (Nishiyama et al., 2006). Further past the threshold of photoinhibition, this excess oxidative stress can cause cell death. Finally, cells must preserve or repair their genomes, which are vulnerable to ultraviolet-induced damage (Llabrés and Agustí, 2006; Agustí and Llabrés, 2007).

Photosynthetic organisms have an equally great variety of mechanisms to cope with stress. First, they can adjust their photosynthetic absorption cross section, so as to capture fewer photons (Huner et al., 1998). Second, the photosystem or chlorophyll will re-emit a larger fraction of absorbed light, increasing its fluorescence. Third, what energy is absorbed and not fluoresced can be non-photochemically quenched (NPQ) and released as heat, for example by the interconversion of xanthophyll pigments, (Falkowski and Raven, 2007). Fourth, electron flow from PSII may be redirected cyclically through plastoquinol terminal oxidase (PTOX), instead of proceeding to photosystem I (Bailey et al., 2008). *Prochlorococcus* employs each of these mechanisms. Both HL and LL ecotypes have some capacity for NPQ, but it is significantly greater in HL (Bailey et al., 2005). Some (but not all) *Prochlorococcus* genomes include a photolyase that repairs UV-induced DNA damage (Coleman and Chisholm, 2007; Osburne et al., 2010). And *Prochlorococcus* has been shown to cycle electrons in a PTOX-like pathway (Mackey et al., 2008).

However, these mechanisms are not likely to be shared across all *Prochlorococcus* ecotypes. It has previously been shown that LL and HL *Prochlorococcus* have differing tolerances not just for sustained growth at particular light intensities (Moore and Chisholm, 1999), but also for short-

term, higher light intensities (Six et al., 2007). One LL clade, however, is an exception: the *Prochlorococcus* NATL ecotype (eNATL) survives mixing events in the wild, which may reflect its ability to survive light shocks in the lab (Appendix C). In doing so, HL and LL strains appear to respond in the same way initially: a decline in *in vivo* chlorophyll fluorescence during the shock, but only HL and eNATL cells recover and return to growth afterward.

If both HL and LL cells moderate their *in vivo* chlorophyll autofluorescence in a light shock and recover it shortly thereafter (Appendix C), the question remains: what genetic features, possessed by eNATL and presumably absent from other LL cells, are responsible for this difference in outcome?

From the sequenced genomes, one possibility stands out: the two sequenced eNATL genomes each contain more than 40 copies of high-light inducible (*hli*) genes (Coleman and Chisholm, 2007). These genes, whose products are known as high light-inducible proteins (HLIPs) or small cab-like proteins (SCPs), have been shown to be upregulated in high light in *Prochlorococcus* (Steglich et al., 2006). They are also upregulated by stress conditions such as nitrogen starvation (Tolonen et al., 2006), phage infection (Lindell et al., 2007, Appendix A), iron stress (Thompson et al., 2011), and carbon limitation (Bagby, 2009). In *Synechocystis*, HLIPs have been shown to be essential for the tolerance of high light shocks (He et al., 2001). However, the mechanism by which they help is unclear: they have been suggested to help shed excess light energy as heat (Havaux et al., 2003), or possibly to bind chlorophyll as it is released from a damaged photosystem II during its repair cycle (Vavilin et al., 2007; Nixon et al., 2010). Further, it has not been shown whether the extra *hli* copies in eNATL confer the observed tolerance for high light. They could alternatively be a mechanism for dealing with other, non-light related stresses. They could even have accumulated in such large numbers as essentially an accident, although this is unlikely given their consistent appearance in islands from a diverse sampling of eNATL cells (chapter 4). It is also unknown what the differences are between eNATL's high light tolerance and that of true HL ecotypes. Do HL cells have a greater tolerance, or do eNATL benefit from their having even more *hli* copies than HL do? Do these two different ecotypes use the same mechanisms to cope with light shocks, the only difference being one of magnitude, or are their different outcomes due to qualitative differences in their response strategies?

To begin to address these questions, I observed physiological properties of *Prochlorococcus* cells exposed briefly to sharp increases in light intensity. We measured *in vivo* chlorophyll fluorescence, both per culture volume and per cell, along with variable fluorescence to evaluate the health of the photosystem. Because the time scale of these physiological decisions was initially unknown, we

performed experiments on time scales ranging from minutes to hours of light shock. Exploration of different timescales showed us surprising differences in the timing of responses between LL cells and HL/eNATL cells. Some of these differences may be a consequence of greater HLIP expression in HL/eNATL. They may also be adaptations in their own rights, with an equal or greater role, compared to that of HLIPs, in determining overall cell survival.

3.2 Methods

3.2.1 *Prochlorococcus* isolates

Four isolates were tested in the experiments described below: NATL2Aax (LL but light shock-tolerant), MED4ax (HL), SS120 (LL), and MIT9313ax (LL). We began with SS120 as a LL point of comparison with NATL2Aax. However, SS120 is not axenic, a possible complication as heterotrophs in co-culture may alleviate oxidative stress in *Prochlorococcus* (Morris et al., 2008; Sher et al., 2011; Morris et al., 2011). When MIT9313ax (MIT9313) became available (Sher et al., 2011), I repeated the long-term recovery experiment (35 to 400 μmol quanta$\cdot$ $m^{-2}s^{-1}$), obtaining results similar to those of SS120, and proceeded with MIT9313ax in subsequent experiments.

3.2.2 Light shocks and bulk culture *in vivo* chlorophyll fluorescence

Cultures were grown in a temperature-controlled incubator, in constant light in 13 mm tubes with Pro99 media (Moore et al., 2007). Bulk chlorophyll autofluorescence was monitored in a 10-AU fluorometer (Turner Designs), with 340-500 nm excitation and 680 nm emission filters. Growth light was provided by Pentron cool white fluorescent lights (Sylvania), covered with mesh screens to achieve the desired light intensity. Light intensity was measured with a QSL-2100 light meter (Biospherical Instruments) and is reported in μmol quanta$\cdot$ $m^{-2}s^{-1}$. Cultures were grown at the starting intensity (10 or 35 μmol quanta$\cdot$ $m^{-2}s^{-1}$) through at least 3 transfers (approximately 3 weeks). For the experiment, cultures were split into 10 identical 5 mL cultures and allowed to grow for three days. Two tubes at a time were transferred to high light (400, 300, or 100 μmol quanta$\cdot$ $m^{-2}s^{-1}$). At 1-hour intervals, additional pairs were transferred. To ensure that the light shock was uninterrupted, no tubes were read or removed from the light during their respective periods at 100 μmol quanta$\cdot$ $m^{-2}s^{-1}$. At hour 4, all tubes were returned to the starting light intensity. Cultures were monitored for an additional three days. At each timepoint (one per day, or once per hour during the shock), bulk fluorescence was measured and samples were frozen for later analysis by

flow cytometry (see below).

1-hour shock experiments were performed similarly. The acclimation intensity was 35 μmol quanta$\cdot$ $m^{-2}s^{-1}$ and the shock intensity was 400 μmol quanta$\cdot$ $m^{-2}s^{-1}$. In this case, replicate pairs of tubes were transferred to high light at 10-minutes intervals for a total of 60 minutes. We used separate cultures for each shock duration out of concern that removing them for reading or sampling at 10-minute intervals would reduce the effectiveness of the light shock. Control tubes remained at 35 μmol quanta$\cdot$ $m^{-2}s^{-1}$ for the duration. Each tube's fluorescence was read, and flow cytometry samples frozen, at the beginning and the end of the 60 minutes' light shock.

3.2.3 FIRe analysis and variable fluorescence

1 L cultures of MED4ax, NATL2Aax, and MIT9313ax were grown at 35 μmol quanta$\cdot$ $m^{-2}s^{-1}$. During mid-log growth, they were split into 6 identical cultures, which were then allowed to grow another 24 hours at the same light intensity. Then, at hour 0, three cultures were moved to 500 μmol quanta$\cdot$ $m^{-2}s^{-1}$ and three remained at 35 μmol quanta$\cdot$ $m^{-2}s^{-1}$.

Fluorescence Induction and Relaxation (FIRe) samples were taken during the 180-minute time-course, simultaneously with RNA sampling. To measure normalized variable fluorescence (F_v/F_m), 200 μL culture was removed, diluted with 1 mL Pro99, and allowed to rest 5 min in darkness. It was then placed in the FIRe instrument (Satlantic) and tested with the included blue LED excitation source and 678 nm filter. The FIRe employs one single-turnover flash (STF), while testing fluorescence with pulses 1 μs apart to capture the single-turnover induction curve, a method similar to fast repetition rate fluorescence (FRRF) (Kolber et al., 1998). STF duration was 150 μs to ensure a plateau at F_m. Single-turnover induction data was analyzed with the Fireworx package (Barnett, 2007). After Barnett's suggestion and our own experience, we excluded the first two datapoints when fitting the induction curve (Appendix F). Additional samples were taken for flow cytometry (see below) and for gene expression (Appendix E).

3.2.4 Flow cytometry and chlorophyll fluorescence per cell

Samples for flow cytometry were preserved as described in Appendix C: 800 μL of culture was mixed with 4 μL 25% gluteraldehyde (0.125% final concentration), held in darkness for 10 minutes, flash frozen in liquid nitrogen, and stored at -80 ° C. Samples were analyzed on an Influx flow cytometer (Becton Dickinson). Scattering and fluorescence per cell are normalized to 2-micron FluorescBrite beads (Polysciences), and cell concentration is calculated based on the measured flow rate on the

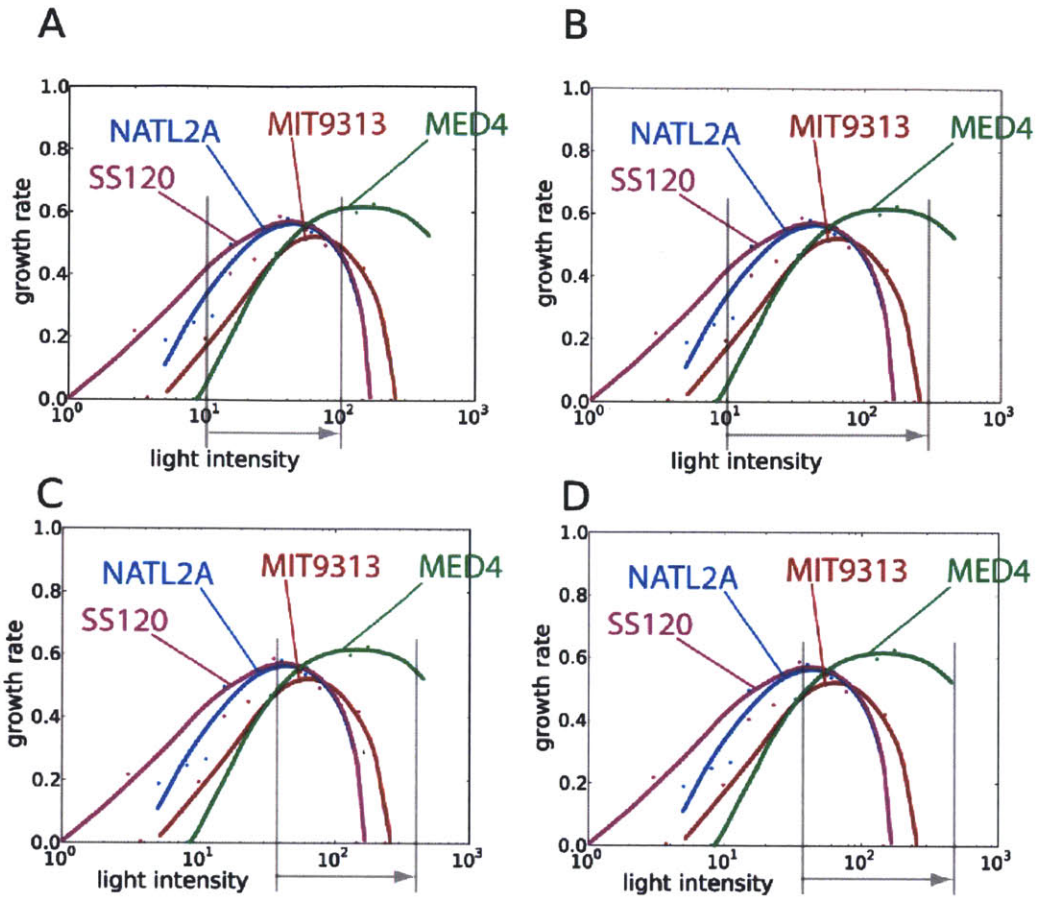


Figure 3-1: Shock intervals used in this study and *Prochlorococcus* growth rate (μ, units of day^{-1}) as a function of light intensity (μmol quanta$\cdot$ m^{-2}s^{-1}) for the four strains tested here. Arrows show the light intensity intervals (A) 10 to 100 μmol quanta$\cdot$ m^{-2}s^{-1} (B) 10 to 300 μmol quanta$\cdot$ m^{-2}s^{-1} (C) 35 to 400 μmol quanta$\cdot$ m^{-2}s^{-1} (D) 35 to 500 μmol quanta$\cdot$ m^{-2}s^{-1}, the acclimation/recovery and shock levels in the various timeseries discussed here. Growth data reproduced from Moore and Chisholm (1999) and Zinser et al. (2007).

flow cytometer.

3.3 Results and Discussion

3.3.1 Light shocks and survival

We have previously reported that MED4ax and NATL2Aax cultures, acclimated to 35 μmol quanta$\cdot$ m^{-2}s^{-1}, survive light shocks lasting 4 hours at 400 μmol quanta$\cdot$ m^{-2}s^{-1} (hereafter written as 35 to 400 μmol quanta$\cdot$ m^{-2}s^{-1}) (Malmstrom et al., 2010, Appendix C). 35 μmol quanta$\cdot$ m^{-2}s^{-1}, the light level of acclimation, is slightly lower than that for fastest growth in a LL isolate, and well below that of MED4 (Fig. 3-1). In (Malmstrom et al., 2010, Appendix C), we considered

the implications of this adaptation for survival during mixing events in the wild. However, this represents only one interval, and a similar fold change could have a different effect at higher or lower light intensities. Shocks of less than 4 hours may also have a different effect. Here, I address those issues by repeating the experiments at different shock intervals (35 to 400 μmol quanta$\cdot$ $m^{-2}s^{-1}$, 10 to 300 μmol quanta$\cdot$ $m^{-2}s^{-1}$ and 10 to 100 μmol quanta$\cdot$ $m^{-2}s^{-1}$), and at intervals of 1, 2, 3 or 4 hours.

After being shifted from 35 to 400 μmol quanta$\cdot$ $m^{-2}s^{-1}$, about a 10-fold increase, SS120 does not recover even after only 1 hour in high light (Fig. 3-2C). MED4ax and NATL2Aax, HL and LL isolates respectively, do recover quickly from this perturbation. Both show a steep decline in fluorescence in the first hour, but recover rapidly to their original fluorescence after returning to intensity to which they are acclimated. Their long-term growth is not inhibited. The same 10 fold increase in light intensity, when using cells acclimated to 10 instead of 35 μmol quanta$\cdot$ $m^{-2}s^{-1}$, had little effect on any strains, other than a small decrease in bulk culture fluorescence during the interval of elevated exposure (Fig. 3-2G,H,I). This may indicate that the maximum tolerable dose of PAR is determined partly in comparison to the cell's acclimation, and partly by an absolute limit. Were it strictly a matter of fold change, a 10-fold increase in intensity starting from 10 μmol quanta$\cdot$ $m^{-2}s^{-1}$ would be the same as 35 to 400 μmol quanta$\cdot$ $m^{-2}s^{-1}$, but this is not enough to harm NATL2Aax, or any LL strain.

Cells acclimated to 10 μmol quanta$\cdot$ $m^{-2}s^{-1}$ and given a light shock up to 300 – a 30 fold change – revealed the "tipping point" for NATL2Aax (Fig. 3-2D,E,F). Under these conditions, they behaved more like LL than HL strains in their response to a light shift. This indicates that NATL2Aax, while light shock-tolerant by the standards of LL ecotypes, is not as robust as a true HL isolate. This also indicates the importance of acclimation, as 300 μmol quanta$\cdot$ $m^{-2}s^{-1}$ is less than the light intensity that NATL2Aax survived in the first experiment (starting from 35 μmol quanta$\cdot$ $m^{-2}s^{-1}$). Interestingly, SS120 cells and NATL2Aax cells seem to have similar tolerances to this light intensity when acclimated to 10 μmol quanta$\cdot$ $m^{-2}s^{-1}$(Fig. 3-2F and G), in spite of the previously-observed difference between them when acclimated to 35 μmol quanta$\cdot$ $m^{-2}s^{-1}$.

Figure 3-2: (Opposite page) Short term response and recovery of *Prochlorococcus* strains to light shock of different intensities. Shaded periods indicate illumination at the initial/acclimated light intensity; the white band at 48-52 hours indicates the light stress period. Different cultures were exposed to 0, 1, 2, 3, or 4 hours of high light, the duration of exposure indicated by the color and symbol of the data. The 4-hour-shocked cultures of MED4ax, NATL2Aax, and SS120 (dark blue triangles) were previously reported (Appendix C). Inset tree indicates the four strain's relation to each other on the *Prochlorococcus* tree from chapter 2.

MED4ax (HL)
NATL2ax (LL)
SS120 (LL)
Light shift (acclim. -> shock, µmol quanta m-2s-1)
10 -> 100
10 -> 300
35 -> 400
culture fluorescence (a.u)
hour
A
B
C
D
E
F
G
H
I
shock duration
ctrl
4hr
3hr
2hr
1hr
MED4
NATL2A
SS120
MIT9313

Figure 3-2

Bulk fluorescence is a relative measure of the culture's state but does not discriminate between changes in cell number and changes in fluorescence per cell. Analyzing cells using flow cytometry was used to further examine this. Cell division virtually stops during a light shock for all isolates (Fig. 3-3), but resumes within one day in MED4ax and NATL2Aax (Fig. 3-3A, B). In SS120, however, division does not recover, and cell numbers decline slowly and continuously for the duration of the experiment (Fig. 3-3C). A closer look at the flow cytometric signatures of the cultures helps us see what is happening during and after this shock. The decline in SS120's bulk fluorescence after a light shock (Figure 3-2C) is due to decline in fluorescence per cell (3-3). A closer look at the distribution of fluorescence across SS120 indicates this is not a uniform decrease. Instead, a sub-population transitions into a lower fluorescence (Figure 3-5C). A similar stepwise shift takes some of the population to a smaller cell size, reflected by the cell's forward scatter signal (Fig. 3-5D). NATL2Aax, when exposed to the same shock (35 to 400 μmol quanta$\cdot$ $m^{-2}s^{-1}$), recovered quickly in bulk fluorescence (Fig. 3-2B), and in cell division (Fig. 3-3B). However, when it is exposed to a lethal shock (10 to 300 μmol quanta$\cdot$ $m^{-2}s^{-1}$), its response shows the same bimodal signature (Fig. 3-6). The bimodal distribution is surprising, as all cells have been exposed to the same conditions. That this does not yield a single, normal distribution suggests an irreversible decision, possibly at a determined point in the cell cycle, leads to a qualitative change in the cells' state. If these shifts in fluorescence and size reflect the process of cell death, they would suggest each cell goes through that process suddenly, but determines the timing independently of other cells. Future investigations could determine if this state change is part of the cell cycle and if so, at what point in the cycle it takes place.

In each isolate, the response to high light is characterized in part by a steep decline in bulk fluorescence in the first 4 hours (Figure 3-2). Surprisingly, LL isolates along with HL isolates recover some fluorescence in the hour after a shock, even though their long-term fate has already been determined. The rapid decline in fluorescence could reflect a loss of chlorophyll, but its rapid decline and rapid recovery suggests this is unlikely. It has long been observed that in vivo chlorophyll fluorescence depends not just on chlorophyll concentration but on its availability and connection to the light-harvesting antenna, which can change through a variety of mechanisms (Kiefer, 1973; Macintyre and Cullen, 2005). It appears, then, that *Prochlorococcus* cells, of both HL and LL ecotypes, are able to alter their exposure to light stress through one of these mechanisms.

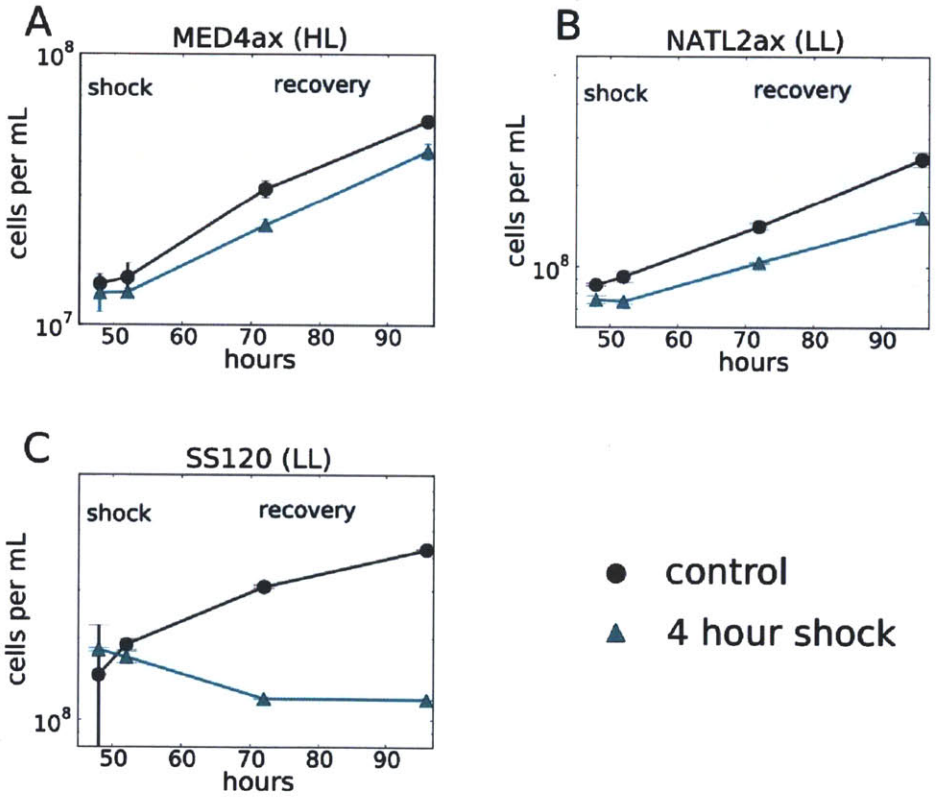


Figure 3-3: Cell counts and survival of *Prochlorococcus* strains after a 4-four, 35 to 400 μmol quanta·$m^{-2}s^{-1}$ exposure, corresponding to Fig. 3-2A,B,C.

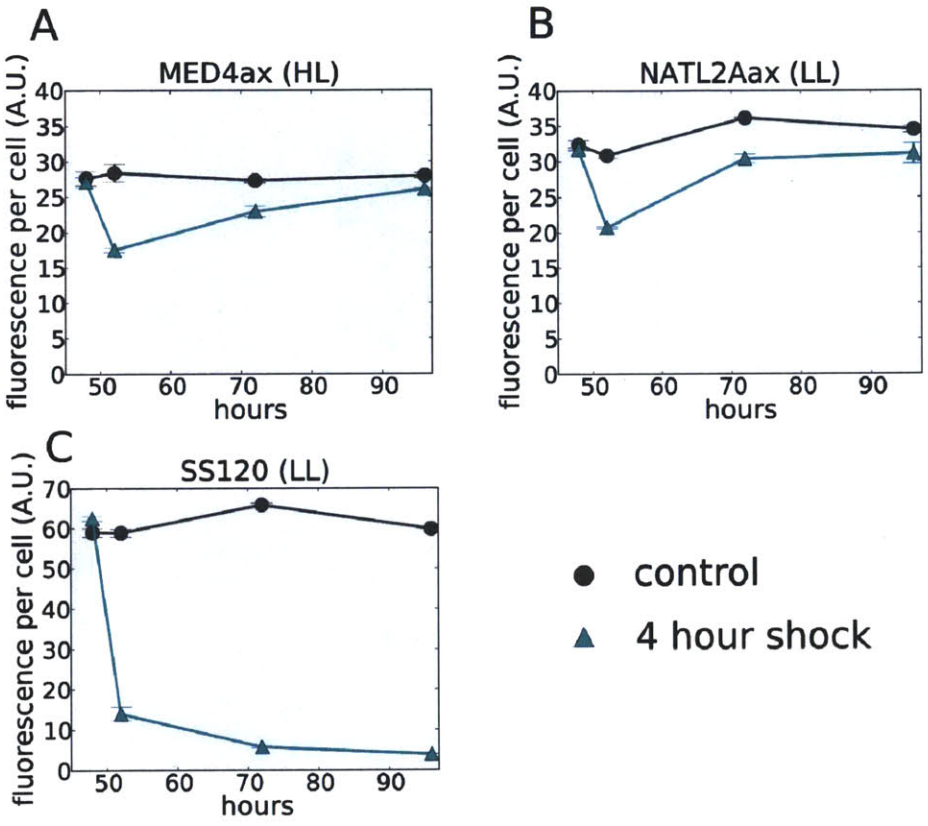


Figure 3-4: Fluorescence per cell (A.U., relative to beads) of *Prochlorococcus* strains after a 4-four, 35 to 400 μmol quanta$\cdot$ $m^{-2}s^{-1}$ exposure, corresponding to Fig. 3-2A,B,C.

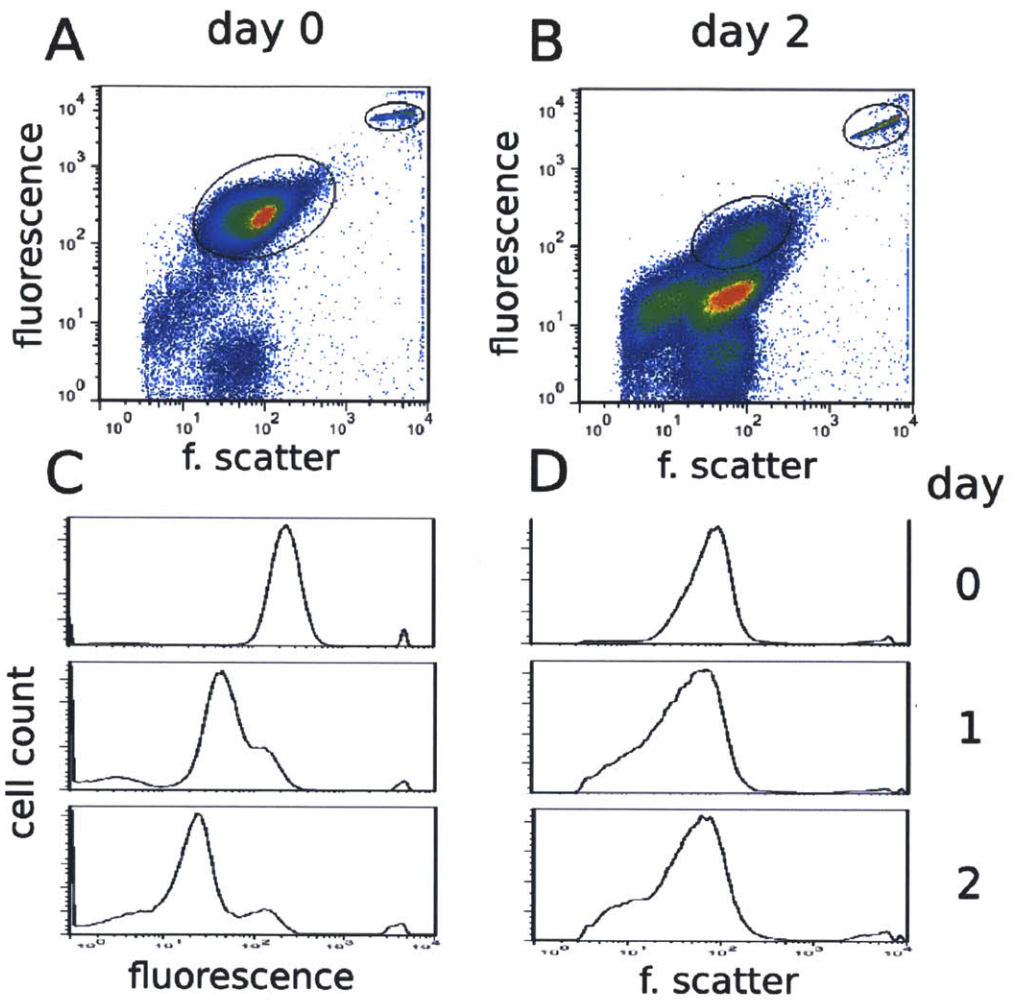


Figure 3-5: Changes in *In vivo* chlorophyll fluorescence per cell and forward scatter (cell size) in an SS120 culture after a 35 to 400 μmol quanta$\cdot$ $m^{-2}s^{-1}$ light shock. (A) and (B) scatterplots at immediately after the shock and 2 days later. (C) Timeseries of fluorescence per cell distribution across the 3-day period. (D) Timeseries of forward scatter (cell size) distribution across the 3-day period.

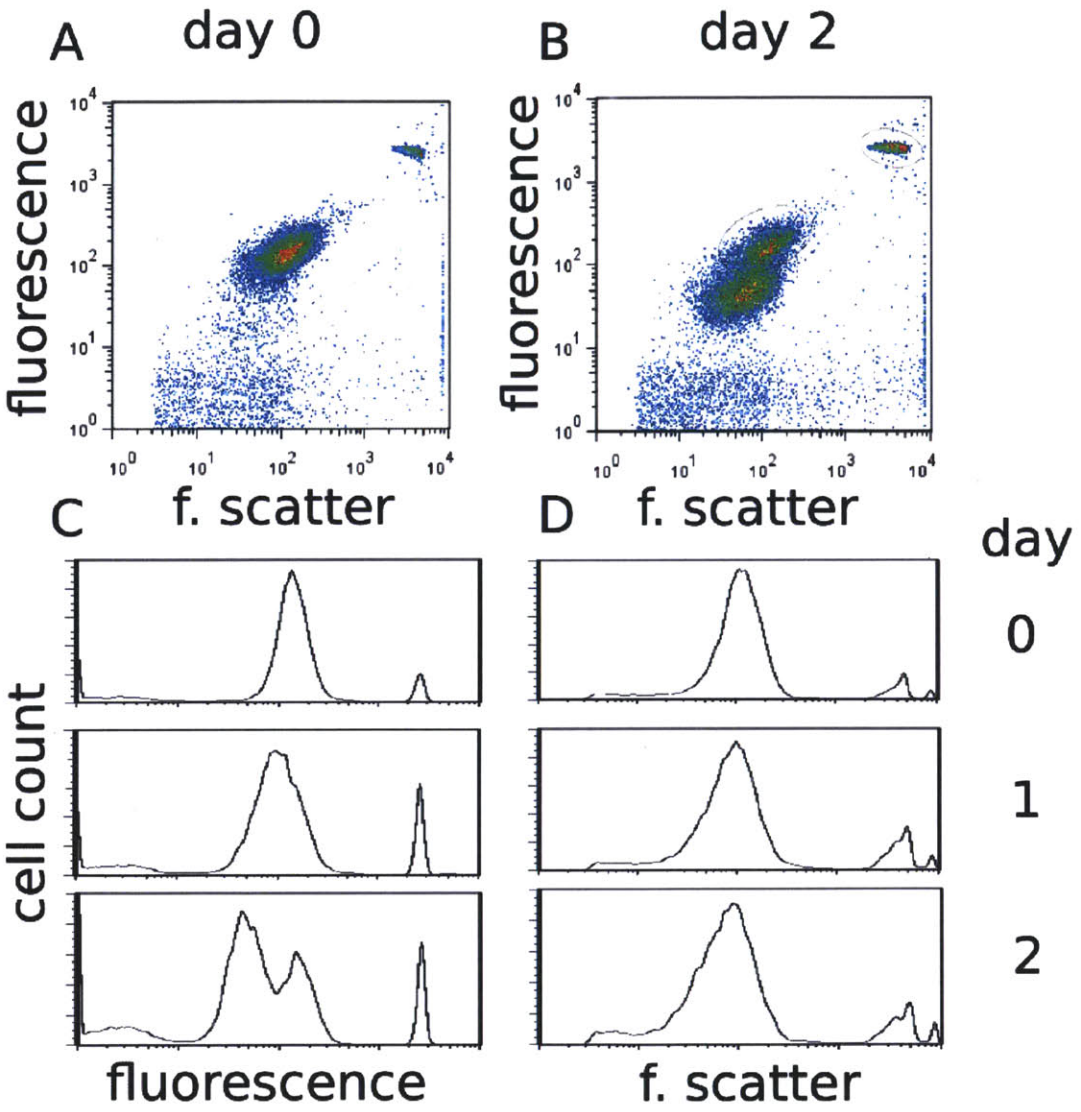


Figure 3-6: Changes in *In vivo* chlorophyll fluorescence per cell and forward scatter (cell size) in a NATL2Aax culture after a 10 to 300 μmol quanta$\cdot$ $m^{-2}s^{-1}$ light shock. (A) and (B) scatterplots at immediately after the shock and 2 days later. (C) Timeseries of fluorescence per cell distribution across the 3-day period. (D) Timeseries of forward scatter (cell size) distribution across the 3-day period.

3.3.2 The short-term response and determination of long term survival

In the experiments depicted in Figure 3-2, samples were collected at 1-hour intervals, as we did not initially know how rapid the cell's first response might be. We noticed, however, that counter to intuition, the initial drop in bulk fluorescence was greater in MED4ax and NATL2Aax than in SS120. MED4ax and NATL2Aax culture autofluorescence declined in the first hour, whereas this decline is delayed until the second hour in SS120. To explore this further, we tested a short-term timeseries, putting cultures into high light at 10-minute intervals for up to one hour of light shock (Figure 3-7). Here, we used both MIT9313ax and SS120 as LL points of comparison with NATL2Aax. The resulting contrast between NATL2Aax and other LL cultures was dramatic: NATL2Aax and MED4 bulk chlorophyll autofluorescence declines in the first 10-20 minutes (Fig. 3-7A and B), which is not the case for SS120 or MIT9313ax. As discussed above, the decline in fluorescence could reflect a contraction of the light-harvesting apparatus or a temporary sequestration of chlorophyll. In the case of SS120 the experiment confirms the previous finding: there is no change in the first hour. MIT9313ax fluorescence even increases slightly in the first hour. This increase could represent the shedding of excess energy by the heavily stressed photosystem.

It seems unlikely that the initial (less than 1 hour) decrease in bulk fluorescence of the NATL2Aax and MED4ax cultures would be solely due to photodamage, as they are better able to cope in the long term than are the other LL isolates. The same amount of chlorophyll can exhibit lower *in vivo* fluorescence for a number of reasons, such as a change in size of the light-harvesting complex (Huner et al., 1998), non-photochemical quenching (Dandonneau and Neveux, 1997) changes in membrane integrity (Loftus and Seliger, 1975), or, in eukaryotes (much larger than *Prochlorococcus*), reorganization of chloroplasts within the cell as a photoprotective measure (Kiefer, 1973). Non-photochemical quenching of fluorescence (NPQ) has been observed in *Prochlorococcus*, and has been seen to be greater in HL strains than in LL (Bailey et al., 2005), and would be a very likely explanation here. No particular gene has been proven to be responsible for the difference here, but a *hli* knockout strain of *Synechocystis* may have a lesser capacity to shed excitation energy as heat (Havaux et al., 2003).

On the other hand, the initial climb in MIT9313ax fluorescence is very likely due to a reduced ability to dispose of absorbed energy photochemically, either due to the complete reduction of the plastoquinone Q_a or damage elsewhere in the photosystem. This would leave excess energy with no release except through fluorescence. While some of the long-term decline of *in vivo* fluorescence

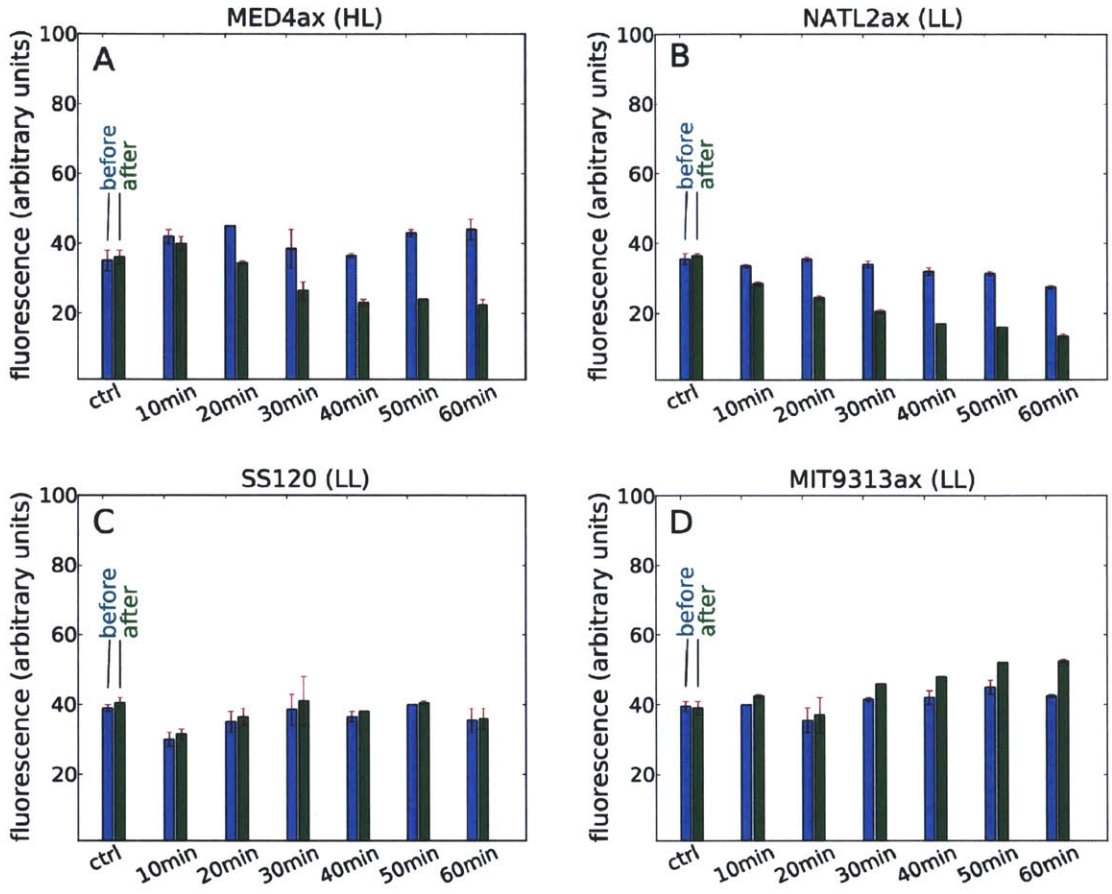


Figure 3-7: Whole culture chlorophyll autofluorescence declines within 10-20 minutes of a 35 to 400 μmol quanta$\cdot$ $m^{-2}s^{-1}$ shift in NATL and HL cells. LL show no decline in the first 1-2 hours.

may be due to a loss of chlorophyll, it appears very unlikely that this is the case in this first hour, especially in the cases of NATL2Aax and MED4ax. Chlorophyll extraction and measurement in solution should confirm this. While overall fluorescence highlights one difference between NATL2Aax and other LL cells, the productivity of the photosystem during these shocks remained unknown. We wondered whether HL and NATL2Aax photosystems remain functional through these shocks and if not, how quickly they might recover. Normalized variable fluorescence (F_v/F_m) provides one estimate of maximum photosynthetic yield, and is another way to evaluate photosystem health during these events. We tracked F_v/F_m with the FIRe system during a 35 to 400 μmol quanta$\cdot$ $m^{-2}s^{-1}$ light shock experiment (Figure 3-8). At 30-60 minutes, F_v/F_m decreased significantly in every isolate. However, in the next hour a difference emerged, with MED4ax F_v/F_m declining only slightly, MIT9313ax continuing a rapid decline, and NATL2Aax in between. The rapid and more severe decline in MIT9313ax relative to the other two strains contrasts with its rising bulk fluores-

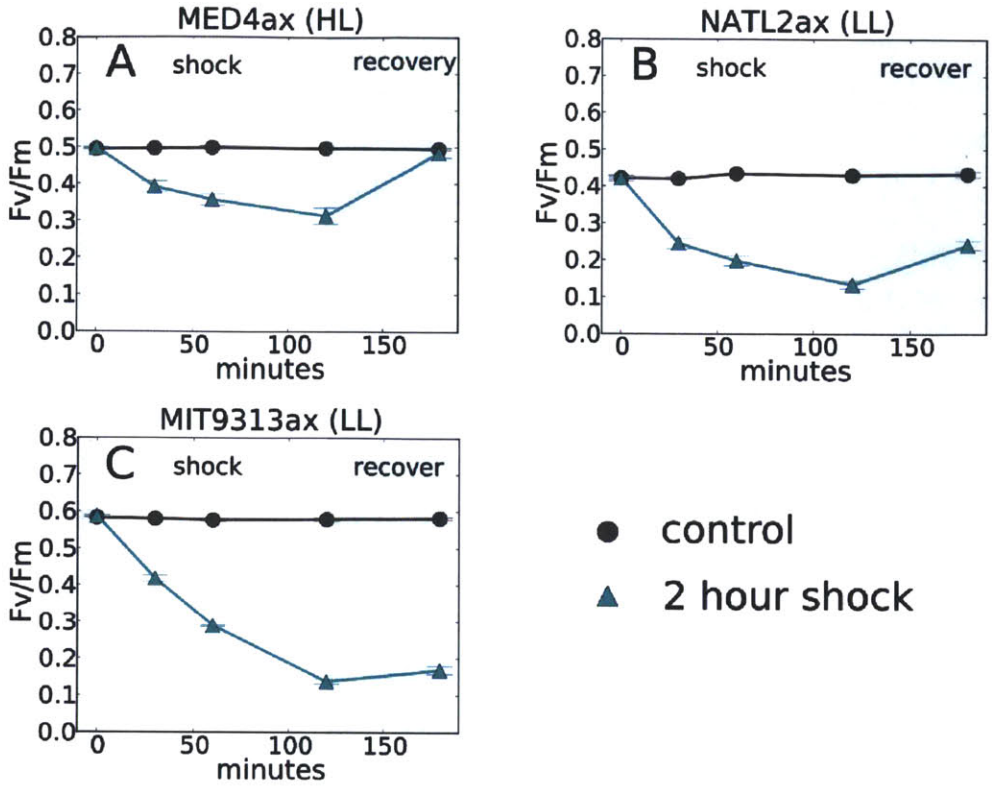


Figure 3-8: Variable fluorescence (F_v/F_m) as a function of time in light shocked cultures. Cultures were acclimated to 35 μmol quanta$\cdot$ $m^{-2}s^{-1}$ and shifted to 500 μmol quanta$\cdot$ $m^{-2}s^{-1}$ for 2 hours, then returned to 35 μmol quanta$\cdot$ $m^{-2}s^{-1}$ and allowed to recover for an hour. (A) MED4ax, (B) NATL2Aax, (C) MIT9313ax.

cence (Fig. 3-7D), which adds weight to the possibility that its photosystem is shedding the light energy it cannot use at this point. Even more dramatic is the difference during recovery: MED4ax, after 60 minutes recovery at the light level to which it was acclimated, fully recovered in F_v/F_m. MIT9313ax showed only a small recovery. NATL2Aax again displayed an intermediate behavior, suggesting, as the survival experiments do, that its tolerance to light shocks falls between that of a true HL isolate and other LL isolates.

3.4 Conclusions and Future Directions

Having previously shown that NATL2Aax survives light shocks that other LL-adapted isolates do not (Appendix C), we sought in this investigation to learn the details behind that difference. Especially interesting were the short-term responses. We suspected that the difference in gene content, particularly in the number of *hli* genes, would be primarily responsible for any difference between eNATL and other LL isolates (Coleman and Chisholm, 2007).

The change in bulk fluorescence and variable fluorescence in MIT9313ax and SS120, and especially their failure to recover, suggest some damage has been done to the cells. However, the nature of that damage, and its extent in MED4ax and NATL2Aax, is unknown. The rapid recovery of NATL2Aax and MED4ax could be explained by the speed of the PsbA replacement cycle, as both grow normally after overnight recovery. In the case of MIT9313ax and SS120, the situation may be more a case of chronic photoinhibition, damage to photosystem I or damage to photosystem II beyond the PsbA protein, and thus beyond the reach of the specialized PsbA degradation and replacement mechanism (Nixon et al., 2010).

Altogether, a picture emerges of SS120 and especially MIT9313 as having a limited response to high light specifically, instead responding only after an hour or more after oxidative stress accumulates. It remains to be seen whether differences in gene content, or in the regulation of what genes are present, is the greater influence on the final outcome in this case. Genome sequencing has allowed us to study gene content in detail, and certainly those dramatic differences must be important in some cases. However, comparative gene expression studies have lagged, partly due to the expense of custom microarrays. The availability of inexpensive RNA sequencing should reverse that trend in the near future, allowing more careful study of these two variables' relative importance (Marioni et al., 2008). If, as preliminary data suggest (Appendix E), we genes shared by NATL2Aax and MIT9313ax but upregulated only in the former, that would argue that gene regulation is equally or more important in determining the two isolates' different outcomes.

The best evidence would be to measure the survival of *hli* knockouts of HL or eNATL isolates. Unfortunately, knockouts are not available to *Prochlorococcus* research at this time. However, further exploration of the uncultured *Prochlorococcus* population, particular by single-cell sequencing, will demonstrate how common the many-*hli* genotype is in eNATL or other *Prochlorococcus* clades, possibly adding further evidence that they are essential to survival in the niche eNATL occupies.

Chapter 4

Chromosomal organization of *Prochlorococcus* genes encoding high light inducible proteins (HLIPs): Insights through metagenomics

Gregory C. Kettler, Steven Biller, Jessie Thompson, Maureen Coleman, and Sallie W. Chisholm

Abstract

Light intensity is one of the primary environmental factors to differentiate *Prochlorococcus* ecotypes. The genetic changes that gave rise to the high light-adapted ecotype are thought to have taken place only once. However, another, laterally-transferred gene family, encoding the high light-inducible proteins (HLIPs), may also contribute to light tolerance, especially under short-term high light exposure. Their numbers in the genomes of one ecologically distinct, low light-adapted clade, the eNATL ecotype, could explain its ability to survive such short-term light shocks that other low light-adapted isolates do not. Past work has also shown that HLIPs are encoded by cyanophage genomes and that most *Prochlorococcus* HLIPs resemble these phage genes, not their homologs in freshwater cyanobacteria. This raises the possibility that these phage-like HLIPs could have a different role in *Prochlorococcus* evolution when compared to their freshwater cyanobacteria-like counterparts.

Here, we investigate the evolution and distribution of both types of *Prochlorococcus* HLIPs in the oceans using metagenomic data. By surveying the HLIPs encoded by short, randomly cloned fragments from

ocean surface waters (the Global Ocean Survey), we find that the freshwater cyanobacteria-like HLIPs have a fixed place in the *Prochlorococcus* genome and are rarely seen associated with island or phage-like sequences. Conversely, we find many more phage-like HLIPs and find that they are localized to highly variable genomic islands previously described in *Prochlorococcus*. We also examine the evolution of HLIPs encoded by one eNATL island by examining large-insert clones that represent that island from a broad sample of wild eNATL cells. While some HLIP-encoding genes in this island are well-conserved across the sequences examined, others are inconsistently distributed and are apparently not maintained. This suggests that while some island-encoded eNATL genes are universal fixtures for the ecotype and could help define its niche, the insertions of additional copies provide no further selective advantage. On the basis of these results, eNATL islands can be said to contain some significant ecotype-defining genes, but also "noise" in the form of other, rapidly changing genes that provide relatively little selective advantage.

Author contributions

G.C.K. planned experiments, analyzed data and wrote the manuscript. S.B. and J.T. prepared Illumina libraries for fosmid sequencing. M.C. helped to select fosmids and to plan experiments. S.W.C. helped to plan experiments and to revise the manuscript.

4.1 Introduction

Laterally transferred genes have a significant role in helping microbes cope with environmental stresses (Hacker and Carniel, 2001). One system in which this phenomenon is especially notable, and increasingly well studied, is *Prochlorococcus*, which provides a significant fraction of the photosynthetic activity in the world's oceans (Goericke and Welschmeyer, 1993). Its ability to survive in a wide variety of environments follows from its diversity of ecotypes. The genes or alleles responsible for traits such as high-light (HL) or low-light (LL) adaptation, or differing temperature adaptation between the two major HL clades, have not been definitively identified (Kettler et al., 2007, Chapter 2, Appendix I)). However, variation within those clades appears to be connected to their collections of laterally transferred genes (Coleman et al., 2006). Such genes include adaptations to specific environmental challenges such as nutrient limitation (Martiny et al., 2006), as apparently these selections took place after the division of the *Prochlorococcus* lineage into its major clades. Our ability to measure the relative abundances of particular ecotypes in particular environments, and to consider this in terms of their physiological optima as assayed in the lab, offers a unique advantage in studying how that diversity relates to the environment. Measurements of ecotype

abundance through quantitative PCR have been informative (Johnson et al., 2006), but new advances through metagenomics (Coleman and Chisholm, 2010) and single-cell sequencing (Rodrigue et al., 2009) promise to reveal additional, environmentally significant adaptations within subsets of those major ecotypes, potentially illuminating the roles of those laterally transferred, island-located genes. For example, phosphate acquisition genes such as *phoA*, *phoBR*, and *pstS* have been found in greater numbers in Atlantic metagenomic samples than in Pacific samples, pointing to the role of phosphate as a limiting nutrient in some locations but not in others (Coleman and Chisholm, 2010).

For *Prochlorococcus*, a photosynthetic organism, variation in light intensity is one of the most significant stresses the cell experiences. The most important ecotypic branching point in *Prochlorococcus* seems to be the division between high light-adapted (HL) and low light-adapted (LL) ecotypes, defined by their optimal light intensities for sustained growth, and by the range of light intensities they can tolerate (Moore and Chisholm, 1999). But beyond that adaption, *Prochlorococcus* isolates also have different tolerances for large but brief increases in light intensity (Six et al., 2007; Malmstrom et al., 2010, Appendix C, Chapter 3) .

In cyanobacteria, one gene family appears to have a prominent role in coping with a variety of stresses, but with light stress in particular. Genes encoding high light inducible proteins, or HLIPs, were originally detected in *Synechocystis* (Dolganov et al., 1995). They are so named because they were originally detected as upregulated in high light. HLIPs are small proteins that bear a strong resemblance to both early light inducible proteins (ELIPs) and chlorophyll A/B binding proteins (CABs) in eukaryotic plants. Though HLIPs were originally compared to the three-helix CABs, which are part of the light-harvesting complex in plants, it has since been shown that eukaryotes also encode similar one-helix proteins (OHPs), and that these too are upregulated in high light (Andersson et al., 2003). Though originally named HLIPs, in *Synechocystis* they are also upregulated by a variety of stress conditions, and so they are perhaps more accurately referred to as small CAB-like proteins (SCPs) (Funk and Vermaas, 1999).

The physiological role of HLIPs/SCPs has been difficult to define, but they clearly have a role in light resistance. In *Synechocystis* PCC 6803, knocking out the HLIP-encoding genes (*hli* genes) eliminates the ability to survive high light shocks (He et al., 2001). Further analysis of the deletion strain showed a hyperactive response in terms of pigmentation changes, compared to wild type, which perhaps reflects a more general response to an unmitigated source of reducing power (Havaux et al., 2003). This has been suggested to point to a role for HLIPs in removing that energy, by

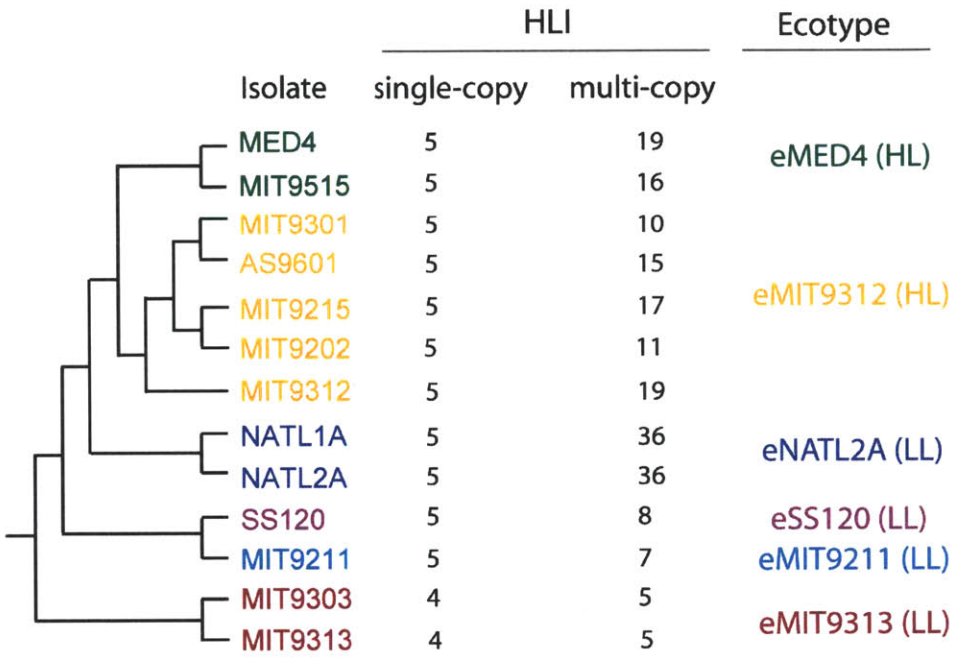


Figure 4-1: Whole genome tree of *Prochlorococcus* as reported by (Kettler et al., 2007, Chapter 2). MIT9202 has been sequenced and added since that publication (Thompson et al., 2011). The number of *hli* genes is as reported by (Coleman and Chisholm, 2007), except for MIT9202 which was analyzed here. The numbers of single-copy, freshwater cyanobacteria-like and multi-copy, phage-like copies as described by (Lindell et al., 2004) are reported separately. The tree is divided into ecotypes, phylogenetically and ecologically distinct populations (Rocap et al., 2002) that can counted in the wild with the use of ecotype-specific quantitative PCR (Johnson et al., 2006).

dissipating it as heat. Further supporting this, the *hli* knockout strain of *Synechocystis* produces more oxygen when exposed to high light than does the wild type (Jantaro et al., 2006). The *hli* knockout is also highly sensitive to excess iron which, as sequestration of free iron is recognized as an antioxidant strategy, suggests a general connection to oxidative stress (Jantaro et al., 2006; Wai et al., 1996). Further, overexpression of bacterioferritin and overproduction of carotenoids (by knockout of the regulatory gene *pfsR*) apparently alleviates oxidative stress and rescues both phenotypes: light shock sensitivity and iron sensitivity (Jantaro et al., 2006).

The HLIPs' exact binding site and partners are also unclear. A number of studies suggest localization to PSII (Yao et al., 2007; Kufryk et al., 2008; Promnares et al., 2006). Others have detected their association with PSI (Wang et al., 2008). More recently, and specifically, it is suggested they bind chlorophyll to prevent its degradation during the PsbA replacement cycle (Vavilin et al., 2007; Nixon et al., 2010; Storm et al., 2008).

However, the accumulating evidence argues against treating all HLIPs as a homogeneous group, especially in the marine cyanobacteria. As the first marine cyanobacterial genomes became available, it quickly became apparent that there are multiple lineages of these genes (Bhaya et al., 2002). In *Prochlorococcus*, the HLIPs proved to be one of the most dynamic gene families. They appear in every sequenced *Prochlorococcus* genome with between 9 and 41 copies (Figure 4-1) (Coleman and Chisholm, 2007), their numbers varying significantly even between closely related isolates, and often located in highly variable islands, suggesting they are recent lateral gene transfers (Coleman et al., 2006). After the sequencing of the first three *Prochlorococcus* genomes, plus three *Prochlorococcus* phages, it was shown that cyanophages carry their own copies of *hli* genes, and that these copies in fact form a phylogenetic cluster with a subset of the *Prochlorococcus* copies (Lindell et al., 2004). Later analysis of additional *Prochlorococcus* and phage genomes suggest that genetic transfers between *Prochlorococcus* and phage are ongoing, and that *hli* genes are particularly prone to such transfers (Sullivan et al., 2005; Thompson et al., In press).

The experimental results discussed above must be considered in light of the HLIP phylogeny in *Prochlorococcus*. The knockout and binding partner studies were carried out in *Synechocystis*, which possesses no multi-copy *hli* genes at all. This leaves the possibility that the multi-copy *hli*s have functions or binding partners different from those being identified in *Synechocystis*. Such a possibility has been proposed before, in light of their sharing a conserved C-terminal motif (TGQIIPGxF) that is always absent from the freshwater cyanobacteria-like (Bhaya et al., 2002).

Phage-like *hli* copies concentrate in highly variable genomic islands in the sequenced *Prochlorococcus*

genomes (Coleman et al., 2006; Kettler et al., 2007, Chapter 2). Indeed, as the concepts of the flexible genome and core genome were later applied to *Prochlorococcus*, it becomes apparent that the notion of "multi-copy, phage-like genes" (Lindell et al., 2004) is a specific instance of what we now call the flexible genome (Kettler et al., 2007, Chapter 2). It also becomes apparent that, besides being single-copy and freshwater cyanobacteria-like, the remaining *hli* genes are conserved as part of the core genome (in the case of 4 of them), or nearly so (a fifth is absent only from MIT9313 and MIT9303, out of all sequenced genomes) (Kettler et al., 2007, Chapter 2). We refer to freshwater cyanobacteria-like *hli* genes as "core *hli*s" here.

The role of *hli*s as stress response genes in *Prochlorococcus* is well established: besides light (Steglich et al., 2006), they are upregulated by stress conditions such as nitrogen starvation (Tolonen et al., 2006), phage infection (Lindell et al., 2007, Appendix A), and iron starvation (Thompson et al., 2011). However, that stress response role may be exclusive to the multi-copy *hli*s. The core *hli*s, on the other hand, were not upregulated in any of the stress response experiments mentioned above (Table 4.1). Curiously, one core *hli* (PMM1482) was actually seen to be downregulated in a carbon starvation experiment (Bagby, 2009).

The differing responses of core and phage-like *hli* copies might point to a division of labor in which core *hli*s are housekeepers and the phage-like, multi-copy *hli*s, concentrated in islands, respond only to extreme stress. Observation of their expression in MED4 growing in a synchronized day-night cycle suggests otherwise, however. Under these ideal growth conditions, there no pattern distinguishes the expression of all phage-like *hli* copies from all core copies. Instead, subsets of both groups are up- or down-regulated at specific points in *Prochlorococcus* MED4's cell cycle (Zinser et al., 2009). As core and multi-copy HLIPs in proximity on the genome may be co-expressed, it is possible that these diel expression patterns are driven more by chromosomal conformation than by particular transcription factors (Jeong et al., 2004; Peter et al., 2004; Willenbrock and Ussery, 2004; Balke and Gralla, 1987; Dorman, 1996). Taken together, the diel experiment and the stress experiments suggest a "double duty" for multi-copy HLIPs, as they are expressed both on a regular cycle and in stress responses. Core HLIPs, not being upregulated during sudden stresses, seem only to have the cell cycle-linked role.

If it is the case that island-borne *hli* genes are general stress response genes, their distribution across the *Prochlorococcus* lineage is curious. In general, high light-adapted *Prochlorococcus* possess more copies per genome than do low light-adapted isolates. However, the two sequenced low light-adapted eNATL genomes carry more copies than any other sequenced *Prochlorococcus* genome

(Coleman and Chisholm, 2007). This may assign them a niche as a transitional low light-adapted ecotype, as their tolerance for brief exposure to high light is greater than that of other low light-adapted cells (Chapter 3 and (Appendix C). Alternatively, the extra copies may be preserved as protection against other forms of oxidative stress. Furthermore, it is clear that large numbers of HLIPs are not the only mechanism for light resistance. The true high light-adapted *Prochlorococcus* such as MED4 tolerate light shocks well beyond what eNATL isolates do, in spite of possessing a smaller number of *hli* copies (Chapter 3).

These observations were made possible through a collection of 13 cultured and assembled *Prochlorococcus* genomes, but it is still unknown if they are universal in wild *Prochlorococcus* populations. In particular, we wondered whether the core HLIPs ever appear in islands or in association with island genes, and conversely, whether multi-copy *hli*s ever appear outside of islands in a broader sample of *Prochlorococcus* genomes. To explore this, we examined the Global Ocean Survey dataset (Rusch et al., 2007), which includes many *Prochlorococcus hli* genes. This allows us to address the questions: are the multi-copy *hli* genes confined to islands, as is the case in cultured *Prochlorococcus*? Similarly, are single-copy *hli* genes only found in consistent locations, without being subject to recent recombination events?

			Stress Condition				
Name	Locus	Core?	Light	Nitrogen	Phage	Iron	Carbon
hli01	PMM0093	yes					
hli02	PMM0064	yes					
hli03	PMM1482	yes					-
hli13	PMM1317	yes					
hli20	PMM0471	yes					
hli04	PMM1118	no	+			+	+
hli05	PMM1404	no	+			+	+
hli06-09 hli16-19	PMM0815-0818 PMM1396-1399	no	+		+	+	+
hli10	PMM1390	no		+			
hli11	PMM1385	no	+				+
hli12	PMM1384	no	+				+
hli14	PMM1135	no	+		+	+	+
hli15	PMM1128	no	+	+			
hli21	PMM0690	no	+	+			+
hli22	PMM0689	no	+	+			+

Table 4.1: Differential expression of *hli* genes in *Prochlorococcus* MED4 under a variety of conditions. +: significantly upregulated; -: significantly downregulated (*hli03*, one condition only). Adapted from (Steglich et al., 2006; Tolonen et al., 2006; Lindell et al., 2007; Thompson et al., 2011; Bagby, 2009).

We also wondered how *hli* genes are organized in islands from genomes representing a broad

sample of one ecotype, in particular the eNATL ecotype. While field data on relative ecotype abundances under different mixing regimes suggests that some exceptional resistance to changing light intensities is common throughout the clade (Malmstrom et al., 2010, Appendix C), our genomic and physiological observations were carried out on only two very closely related representatives, both from the North Atlantic. This raises questions: if extra *hli* copies are an ecotype-defining genetic feature in eNATL, they would have to be common to all cells in that clade. If that is the case, how are they organized? Did they insert into islands in an eNATL ancestor and remain static since, or did their insertions or deletions continue even as the clade evolved? To expand the eNATL sequence dataset, we sequenced eNATL-like fosmids from the Pacific Hawaii Ocean Timeseries (DeLong et al., 2006). By targeting a specific, *hli*-containing island region, we greatly expanded the coverage of the eNATL clade, discovering how *hli* copy number may vary throughout this part of the tree.

4.2 Methods

4.2.1 Global Ocean Survey HLIPs

To generate our dataset, we used sequences from the GOS Atlantic, Pacific, and Indian Ocean datasets (Rusch et al., 2007). Previous studies of whole genomes took advantage of conserved sequence motifs to identity HLIPs (Bhaya et al., 2002; Lindell et al., 2004; Kettler et al., 2007, Chapter 2). Given the much larger GOS dataset, those motifs offered an unacceptable trade-off between sensitivity and specificity. Therefore, here we identified HLIPs by using a two-tiered approach. First, HLIPs were identified in all of the GOS reads by scanning all open reading frames for the motifs, AExxNGRxAMIGF (at least 7 matches required) or TGQIIPGxF (7 matches). This is a slightly stricter threshold than that used by Lindell et al. (2004). This yielded a set of 6843 possible HLIPs, including some likely false positives (long, unique proteins that may have matched the query by chance) and missing some false negatives (detected in the next step).

Next we used these HLIPs and HMMER to generate Markov models for a more accurate search (Finn et al., 2010). To simplify the dataset, we removed sequences with more than 80% identity using T-Coffee's trim feature (Notredame et al., 2000). We also removed their N-terminal sequences up to the beginning of the HLIP motif, as HLIP N-terminal sequences are divergent and align poorly. The C-terminal sequences were then aligned using MUSCLE (Edgar, 2004). We reviewed the alignment by hand, removed poorly-matched sequences that were unlikely to be true HLIPs,

separated these sequences into five groups, and re-ran the alignment of each group. We also generated a sixth alignment of the *Prochlorococcus* freshwater cyanobacteria-like HLIPS, along with orthologs from freshwater cyanobacteria and related eukaryotic sequences originally identified by Dolganov et al. (1995). These alignments, or representative samples, are depicted in figure 4-2. With HMMER (hmmbuild), we generated models from these six alignments. With these models and an e-value cutoff of 10^{-3}, HMMER (hmmsearch) found 7106 putative HLIPs with few or no false positives. The HMMER method is described in additional detail in Appendix H.

4.2.2 Fosmids

Based on their end sequences, we selected fosmids likely to have originated from *Prochlorococcus*, and in particular from an *hli*-containing genomic island, from a published collection for sequencing and analysis here (DeLong et al., 2006). Fosmid libraries were prepared as described by DeLong et al. (2006) and their ends were sequenced using Sanger technology at the Department of Energy Joint Genome Institute. Fosmid end sequences were aligned against the assembled *Prochlorococcus* genomes, and those that matched well-conserved NATL1A and NATL2A core genome segments, but oriented so that the complete fosmid should include island sequences, were selected for complete sequencing.

Fosmid DNA was isolated after treating library cultures with CopyControl Induction Solution (Epicentre) to maximize fosmid yield relative to *E. coli* genomic DNA. Illumina libraries were prepared as described by Rodrigue et al. (2010) and sequenced on an Illumina Genome Analyzer GAIIx at the MIT Bio Micro Center. Fosmids were assembled into large (10+ kb) fragments with CLC Assembly Cell (CLC bio, Denmark). Fosmid genes were called by CRITICA (Badger and Olsen, 1999). Additionally, all open reading frames were scanned for HLIPs as described above, and all open reading frames were compared to previously annotated *Prochlorococcus* genes. Fosmid HLIPs were classified using the same method as above. They were also aligned with MUSCLE and a tree was constructed with PhyML (Guindon et al., 2010). Here, we report 12 fosmids that match a small, *hli*-rich island in NATL1A and NATL2A. The complete set is described further in Appendix G.

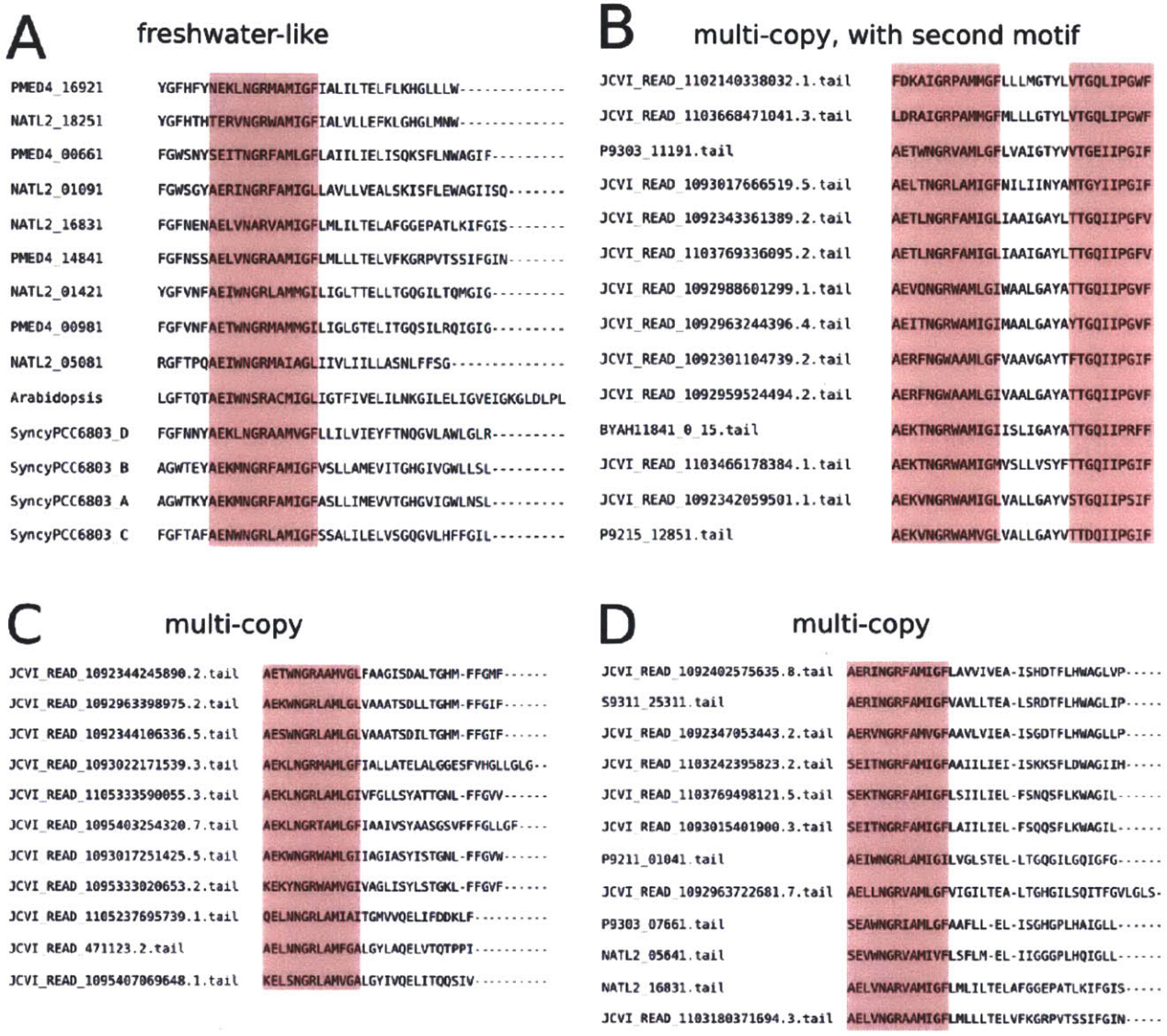


Figure 4-2: Alignments of selected HLIPs from this study. (A) Core HLIPs, their orthologs in *Synechocystis*, and an *Arabidopsis* one-helix protein (OHP) (Andersson et al., 2003). (B) Phage-like HLIPs that include the TGQIIPGxF C-terminal motif. (C-D) Examples of other phage-like HLIPs that do not include the C-terminal motif. Because they are diverse and difficult to align outside of the motif, we break HLIPs into separate groups, of which C and D are examples, here and in using HMMER. Note that the alignments include more sequences than those depicted here. The conserved AExxNGRxAMIGF motif is highlighted, as is the TGQIIPGxF motif that appears in some phage-like HLIPs (Bhaya et al., 2002). All aligned proteins are trimmed of their N-terminal sequences before the region shown here, as those regions vary widely in sequence and length and cannot be aligned.

4.3 Results and discussion

4.3.1 *Prochlorococcus* HLIPs in The Global Ocean Survey Database

The roughly 7000 HLIP copies recruited from GOS reads allow us to explore features of this gene family in wild *Prochlorococcus* populations. We set out to determine the genomic location of the two families of hli genes found in *Prochlorococcus*: those that are core, i.e. shared by all cultured strains, and those that are not. In particular, we wished to know whether core HLIPs have been subject to recombination in the uncultured *Prochlorococcus* population. We also wished to know whether multi-copy HLIPs are truly as island-exclusive as they appear in the 13 genomes sequenced so far.

The core genes are those most similar to the *hli* genes possessed by freshwater cyanobacteria. The phage-like genes belong to a family of hli genes that are also commonly found in phage (Lindell et al., 2004), presumably acquired from host cells, undergoing evolution while resident in phage, and often transferred back to host cells. They are referred to as multi-copy hlis, whereas the core hlis always occur as single copy genes, and are phylogenetically distinct from the phage-like hlis in cyanobacterial genomes. They could also be called "phage-like," as it appears they diverged from the core copies and evolved in phage for an extended time before being re-integrated into the host. From another perspective they would still be called "host-like," when found in the phage genome, but they are referred to as phage-like in discussing their place in *Prochlorococcus* genomes here.

To estimate the locations of *hli* genes in the genomes of uncultured cells, we used the paired end sequence of the GOS read that contained the *hli* homolog, and not the *hli* sequence, as the query sequence when recruiting reads to the genomes. We found that the genomic location of these genes in the GOS reads was consistent with those observed in the cultured strains (Fig. 4-3, Fig. 4-4, supplemental figures). Considering first the core *hli* genes and their locations in the available reference genomes (Figures 4-3C, 4-4C), we see that almost without exception, the 1375 "core" HLIPs identified in the GOS reads are associated with the same genome positions as their homologs in fully assembled genomes. This suggests a much lower degree of recombination involving these genes than other *Prochlorococcus* genes. Conversely, phage-like *hli* copies are almost exclusively associated with sequences matching *Prochlorococcus* islands (Figure 4-3B). The same pattern is observable when querying the low light-adapted, but *hli*-rich, NATL2A genome, although the overall number of hits is lower as the GOS reads come from surface samples where this clade is less abundant (Figure 4-4).

This concentration of a genes of a particular function into islands has been observed before in the case of phosphate metabolism (Martiny et al., 2006; Coleman and Chisholm, 2010) and iron metabolism (Thompson et al., 2011). Also notable is the highly consistent location of the 5 (or 4 in eMIT9313) core *hli* copies, as the copies in GOS are almost always associated with the same sequence neighborhood. Considering the consistent observation of the same set of core *hli* copies in the assembled *Prochlorococcus* genomes, their number and location may have an advantage that leads to their being preserved by selection, unlike the island copies.

We also find that duplication into tandem repeats is a phenomenon exclusive to the island-borne, phage-like *hli* genes. None of the core-classified *hli* genes in GOS can be detected on the same insert as another *hli* gene. Due to their short length and tendency to accumulate in tandem arrays, many are on the same read as at least one other, for a total of 5500 on 3761 reads. The possibility should be considered that the function of phage-like *hli* genes depends on, or is enhanced by being located in tandem arrays, for example by being co-transcribed as operons. If this is not an artifact of their insertion into the genome, they may gain some advantage from being co-transcribed as operons.

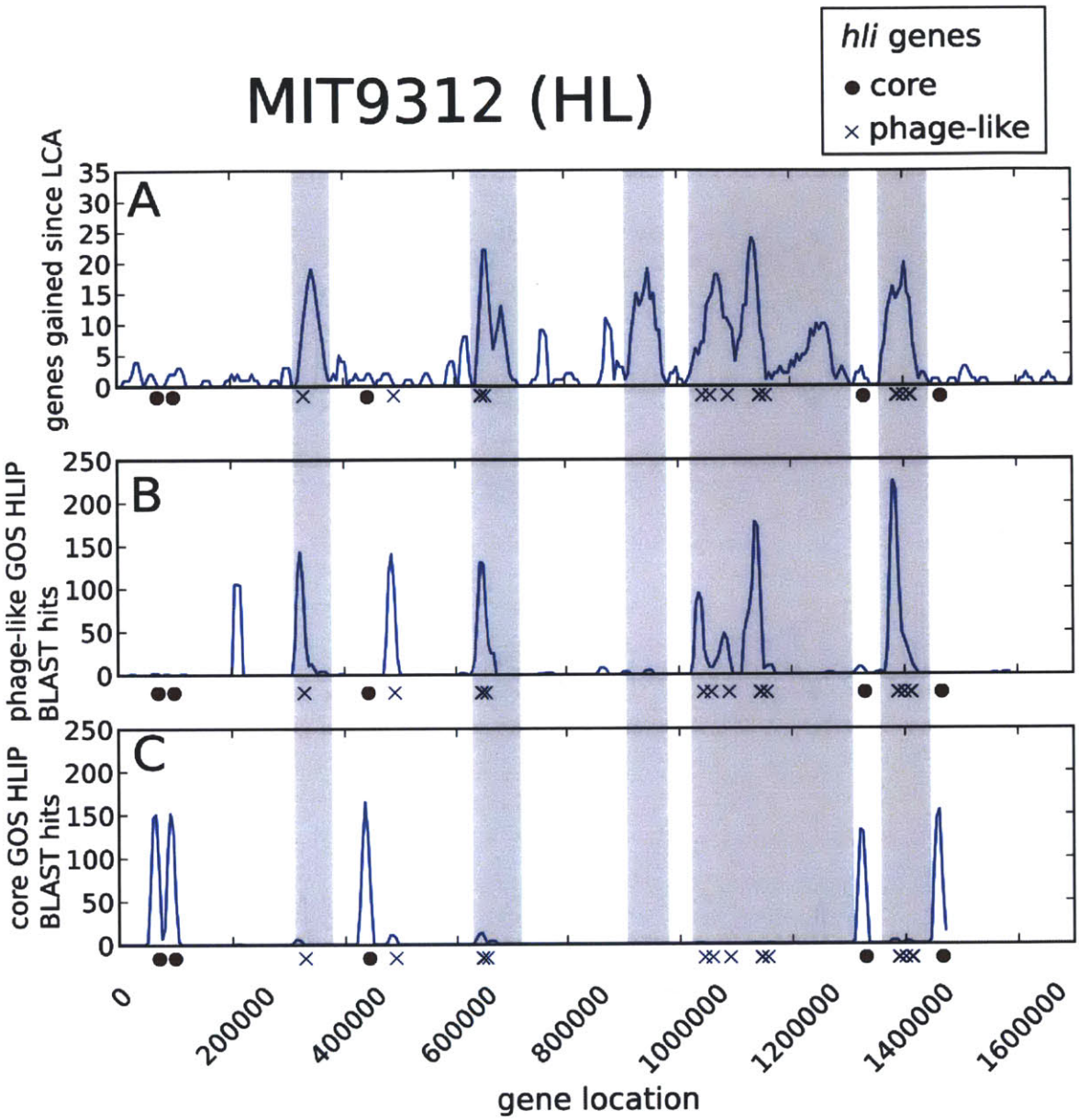


Figure 4-3: Locations of BLAST hits of HLIP-encoding GOS inserts on the high light-adapted *Prochlorococcus* MIT9312 genome. (A) estimates of gene gain locations, providing island locations as in (Kettler et al., 2007, Chapter 2). Locations of the *hli* genes in MIT9312 are along the bottom of each graph, classified as core/freshwater cyanobacteria-like or as phage-like. (B) locations of BLAST hits of GOS inserts encoding phage-like HLIPs. (C) locations of BLAST hits of GOS inserts encoding core-like HLIPs. All plots are sliding windows, window size 15kbp, incremented 5kbp at a time. The shading, for reference, corresponds to the largest peaks (islands) in (A).

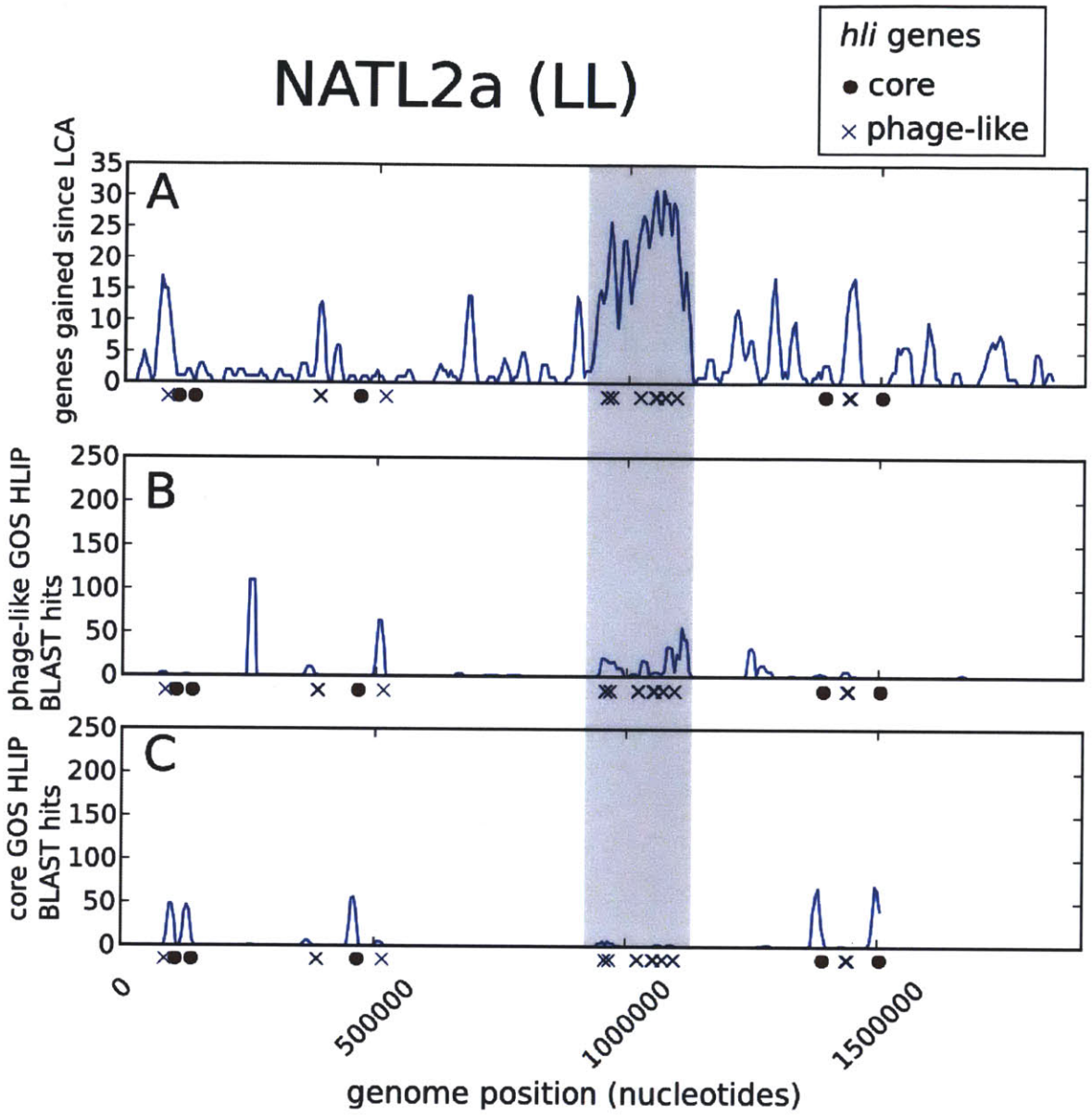


Figure 4-4: Locations of BLAST hits of HLIP-encoding GOS inserts on the low light-adapted *Prochlorococcus* NATL2a genome. Layout is the same as in figure 4-3.

name	% nucleotide match to NATL1A	# *hli* copies	depth (m)	date
(AS9601)	78	3		
FNFS3033	83	4	110	19-Oct-2006
FNFS836	84	4	110	19-Oct-2006
FNFS4955	84	4	110	19-Oct-2006
FNFS7729	86	4	110	19-Oct-2006
BYAH12838	88.2	3	125	9-March-2006
FNFS3863	87	3	110	19-Oct-2006
FNFS4883	87	3	110	19-Oct-2006
FNFS5165	88	5	110	19-Oct-2006
FNFS6078	89	3	110	19-Oct-2006
FNFS7293	84	3	110	19-Oct-2006
BYAH17882	89.9	4*	125	9-March-2006
FNFS3726	96.5	5	110	19-Oct-2006
(NATL2A)	97	5		

Table 4.2: Fosmids reported in this study, and their sample dates and depths. Both samples were taken from HOT station ALOHA: 22° 45' N, 156° 00' W (DeLong et al., 2006). Percent nucleotide match is calculated from the best-matching BLAST (nucleotide) local alignment in the island region in question.

4.3.2 High light inducible genes in fosmids

The two cultured, sequenced representatives of eNATL present a large number of *hli* copies, but are also very similar to each other after being isolated at the same location and date. They raised a number of questions: how much do eNATL cells from other locations differ? Were these extra *hli* copies (all of them in islands) all acquired early in the eNATL lineage, or are they still being frequently gained and lost? If they are still being acquired with any frequency, we also wondered about the origins of the multi-*hli* operons. A single gene might integrate and act as a site for other *hli* genes to recombine, or a whole operon of *hli*s might simply be copied as a unit.

We targeted a 70kb region in the NATL1A and NATL2A genomes, including a 20kb highly variable region between the core genes *murD* and *rpoZ*, which is small enough to be spanned by one fosmid. 24 fosmid contigs had strong matches to this region, and we consider their origins to be eNATL cells as they are more similar to NATL1A or NATL2A than to any other sequenced genome (Table 4.2). The 12 most informative are displayed in figure 4-5. Of particular interest are the tandem *hli* genes that accumulate in all of these genomes. In this region, NATL1A and NATL2A each have 5 *hli* copies, while the fosmids have 3-6 (Fig. 4-6).

It is important to note that this is also an island region in other *Prochlorococcus* genomes, including AS9601 as illustrated (Figure 4-5). This region includes *hli* copies for many of them as well.

All twelve fosmids, and the homologous regions of the NATL1A and NATL2A genomes, share a single array or operon of 3-4 genes (A,D, and C in the detail Figure 4-6), implying it was acquired at some point early in the eNATL lineage and has since been preserved. However, one curious feature is the fourth *hli* gene (E) seen in FNFS3033 and its closest cousins (Fig. 4-6). It is possible that this is one of the repeat expansions that might explain the accumulation of multi-*hli* operons. However, the "extra" *hli* copy (part of clade E) in FNFS4955 bears very little similarity to the other genes on any of these fosmids (Figure 4-7B and Figure 4-6). This suggests that if it was gained in the four fosmids that do possess it, it must be of foreign origin, either from elsewhere in the genome or from a recent lateral gene transfer. Alternatively, the four-gene operon could have been acquired once, and the fourth gene lost in most of the genomes represented here.

These island fragments should represent a broad sample of the entire eNATL lineage (Table 4.2), so it is likely that this operon of 3-4 *hli* copies appears in the majority of uncultured eNATL cells. Next to it, however, it is surprising that the other *hli* copies seen here appear only sporadically. NATL1A and NATL2A, and some of the more closely related fosmids, each contain a pair of *hli* genes, slightly upstream of the trio discussed above (the B clade in Figs. 4-6 and 4-7B). However, most fragments do not, so these genes are either frequently gained or frequently lost. They may provide little selective advantage, a position reinforced by the observation that two clones, BYAH17882 and FNFS4883, each carry an apparently degraded pseudogene copy (Fig. 4-6). Whatever selective advantage that cell gained from this additional *hli* is apparently not enough to maintain it. Instead, eNATL cells could derive much of their stress tolerance from a few functional *hli* operons such as the ADC operon in Fig. 4-6, while additional insertions are then redundant.

The prevailing theory is that *Prochlorococcus* cells recombine foreign genes, possibly from cyanophage, into island regions, and those that provide a selective advantage are preserved (Lindell et al., 2004; Coleman et al., 2006). It also appears, from the number of copies in the eNATL genomes and these fosmids, and from the ability of eNATL cells to survive light exposure in mixing events (Malmstrom et al., 2010, Appendix C), that having extra *hli* copies may offer a selective advantage, relative to other low light-adapted strains. However, it is unclear how active and ongoing this accumulation is today, and what pressures may act against it. One possibility is that the A-D-C operon is functional and beneficial, and therefore an ecotype-defining feature for eNATL, but the other *hli* insertions are less so. This would explain why a BYAH17882 ancestor can still thrive when one copy is degraded.

It is in some ways surprising that these diverse fosmids would have such different arrangements of *hli* copies, as they might be expected to converge to some optimum number. Either these different eNATL cells are each adapted to slightly different environments, or their diversity at this level might be something of an accident. This seemingly chaotic gain and loss of *hli* copies is striking when contrasted with the stability of the core *hli* set, with the same 4-5 genes are present in each of the *Prochlorococcus* genomes, always in the same location.

Figure 4-5

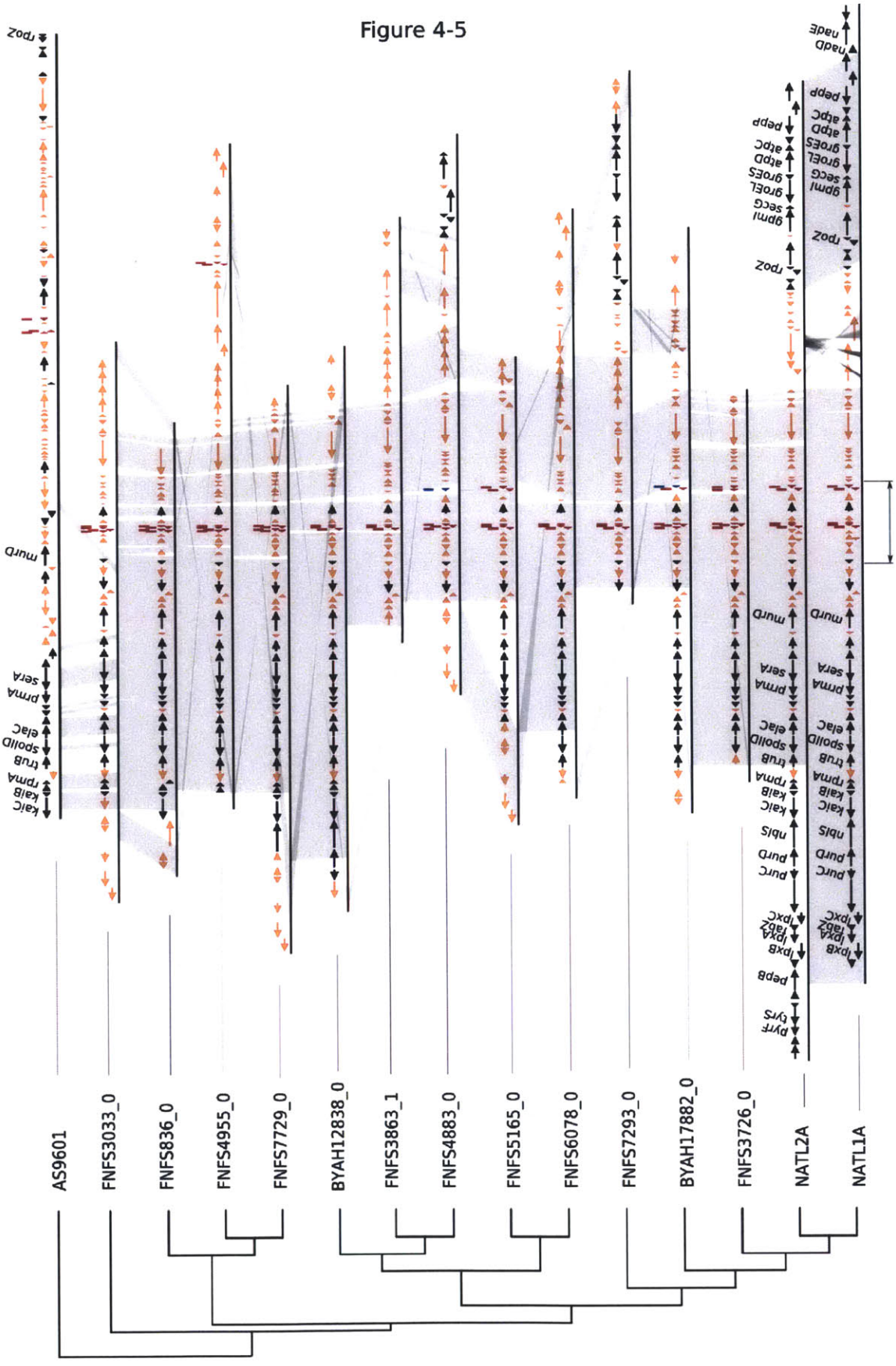

AS9601
FNFS3033_0
FNFS5836_0
FNFS4955_0
FNFS7729_0
BYAH12838_0
FNFS3863_1
FNFS4883_0
FNFS5165_0
FNFS6078_0
FNFS7293_0
BYAH17882_0
FNFS3726_0
NATL2A
NATL1A

Figure 4-5: (Opposite page) Alignments of fosmids against *Prochlorococcus* shows the appearance and disappearance of *hli* repeats. The accompanying tree is based on 9 genes shared by the genomes and all fosmids in the figure. With the exception of FNFS7293, fosmids are vertically arranged in increasing order of identity to NATL1A. Core *Prochlorococcus* are colored black and named where applicable; genes from the flexible genome are orange; *hli* genes are red and are marked with a bar above each; two *hli* pseudogenes are blue. The area of detail in figure 4-6 is marked. The accompanying tree is a PhyML tree of concatenated predicted gene products shared by all fosmids in the figure (Guindon et al., 2010).

4.4 Conclusion

The *hli* genes in *Prochlorococcus* genomes are interesting because they straddle the line between core and flexible genomes. While the homologs of the core *hli* genes in freshwater cyanobacteria are well-studied and responsive to high light, they appear not to be light shock-responsive in *Prochlorococcus* MED4 (Steglich et al., 2006), instead being differentially transcribed across the day-night cycle. If the core *hli* genes lack the regulatory response to stresses that deviate from the day-night cycle, it is possible that their physiological role is specialized such that they would not benefit a stressed cell. The (generally more numerous) multi-copy *hli* family would then be the sole stress response. This is also consistent with other results finding a large portion of stress-response genes are localized to islands, (Coleman et al., 2006).

Furthermore, even among the flexible genome, *hli* genes are unusual. Their tendency to expand into tandem repeats is exceptional even among other island genes. A few possibilities could explain their accumulation into tandem operons. First, gene insertion into islands depends on foreign DNA recombining with similar sequences on the chromosome, so existing *hli* copies could therefore be such recognition sequences. Alternatively, there may be an advantage to their being co-regulated as the same operon. The final possibility might be some local recombination that leads to a repeated *hli* gene, but that is excluded by the apparently foreign origin of the fourth copy in those fosmids that have a four-gene operon.

It is however impossible to know with certainty, by studying a single region from wild cells, what is the true number of *hli* copies in the average uncultured cell. A logical subsequent study would examine other island regions in the same manner. While a small island was selected here, the bulk of the NATL1A and NATL2A *hli* copies appear in another, large island (Figure 4-4), and are thus too far from any core genes to reliably assign a fosmid to a particular ecotype. The alternative to large-insert fosmid clones is to sequence complete single-cell genomes from the ocean, a process

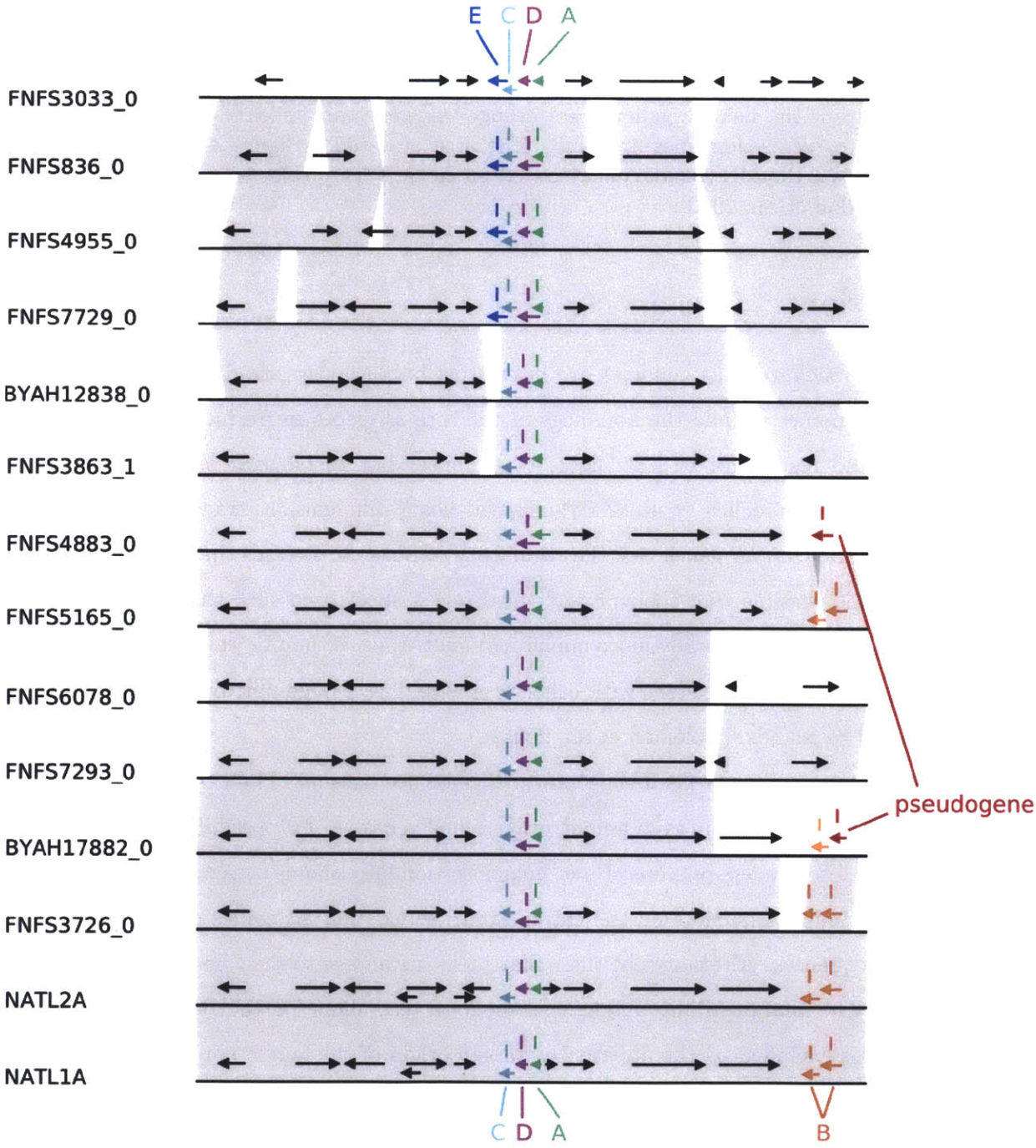


Figure 4-6: Detail of figure 4-5 showing the *hli* tandem array that is common across eNATL. Letters and color codes correspond to the coding in figure 4-7B. Two fosmids, BYAH17882 and FNFS4883, each include a pseudogene resembling a B clade *hli* copy.

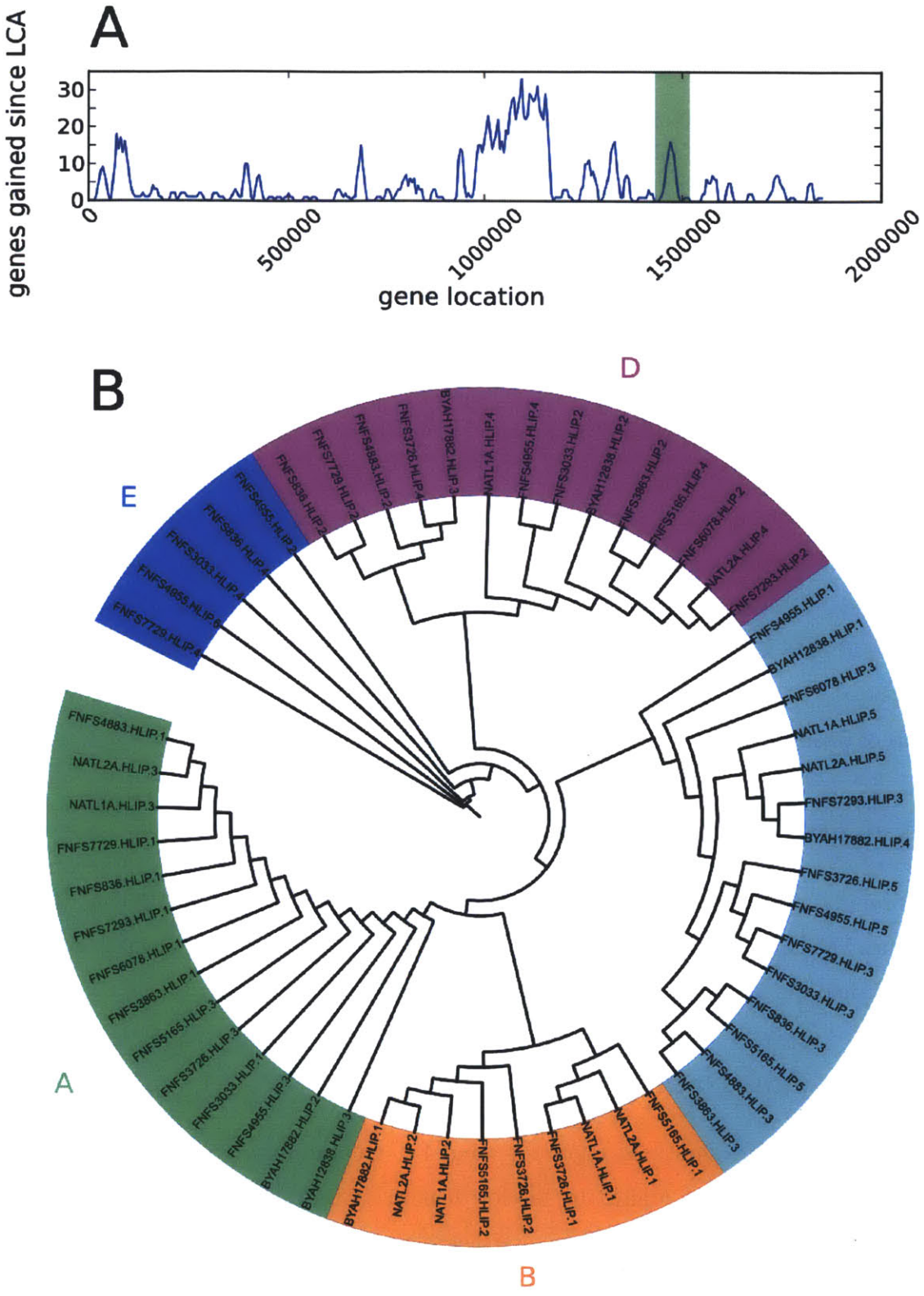


Figure 4-7: (A) The location of figure 4-5 on the overall NATL1A genome is highlighted. Note that it spans a small island, as defined by genes gained since the *Prochlorococcus* last common ancestor, estimated in (Kettler et al., 2007, Chapter 2). (B) Tree of HLIPs encoded by the fosmids in this study. Major clades are color-coded and identified by a letter, both of which correspond to those used in figure 4-6. The tree figure was prepared with Interactive Tree of Life (Letunic and Bork, 2007).

that is now underway and which will be able to confirm the conclusions presented here.

The continuous changes in *hli* copy number, expansion or contraction of tandem *hli* operons, and apparent insertion of new *hli* operons into cloned chromosomal fragments covering the entire eNATL branch may be a model of phage gene integration into host genomes. What is striking here is that, even in the same island, one operon of 3-4 *hli* genes seems to be a universal feature in eNATL, while other *hli* insertions are scattered about the fosmids without sweeping through the population, or even necessarily being maintained where they are inserted. We cannot say, simply, that a cell with more *hli* copies is always better adapted than one with fewer, given the variety seen in the fosmids here. It would therefore be informative to see if those *hli* copies that are not universal across the eNATL clade (B in Fig. 4-6) are upregulated the same way as those that are (A,C,D). This depends on expression data for NATL1A or NATL2A, or alternatively, additional metagenomic sequences to compare against MED4, an isolate for which expression data is most plentiful. Additional insertions may be redundant, or they may be less functional. It may be that, as we sequence additional large inserts or complete genomes from the uncultured population, we will be able to estimate the importance of a particular island gene from the diversity of genomes in which it appears. It may also be that even island genes from the same family do not have the same patterns of distribution.

4.5 Acknowledgements

The authors are grateful to Anne Thompson, Jake Waldbauer, Matt Sullivan, Ed Delong, Tracy Mincer, Yanmei Shi, and Sarah Bagby for field sampling, and to Asuncion Martinez and Jay McCarren for fosmid library preparation.

4.6 Supplemental Data

The supplemental figures are similar to Figures 4-3 and 4-4. GOS reads recruited to the two LL genomes here, like NATL2A, are limited by the low number of LL cells in the surface samples that were sequenced. The pattern in SS120 is nevertheless consistent with those in other genomes (Fig. 4-9). The arrangement of *hli* genes in MIT9313 is dramatically different from those of other genomes and two core *hli* genes may be on the border of an island region, but this is unclear due to the relatively poor definition of eMIT9313 islands in this type of plot.

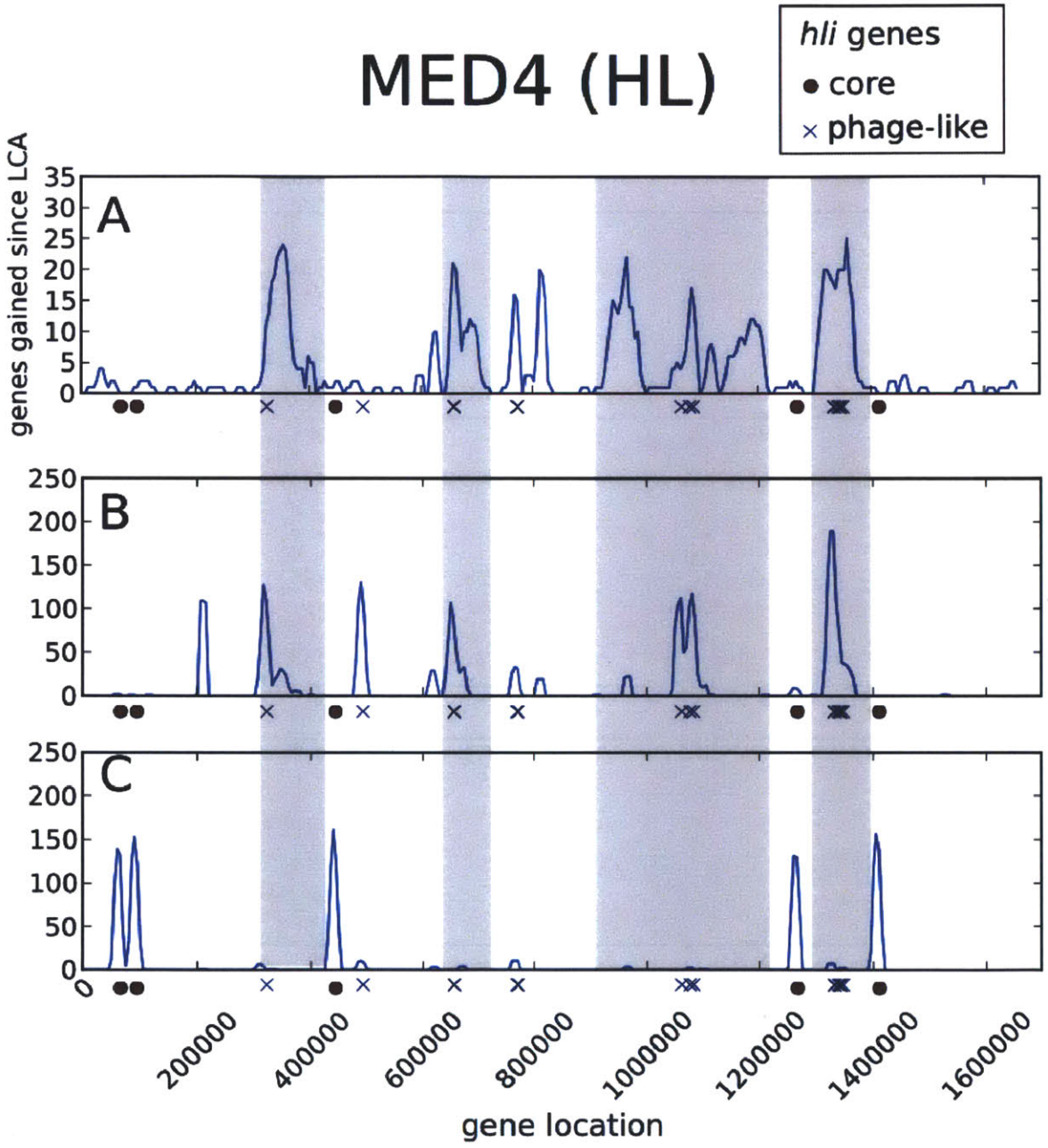


Figure 4-8: Locations of BLAST hits of HLIP-encoding GOS inserts on the low light-adapted *Prochlorococcus* MED4 genome. Layout is the same as in figure 4-3.

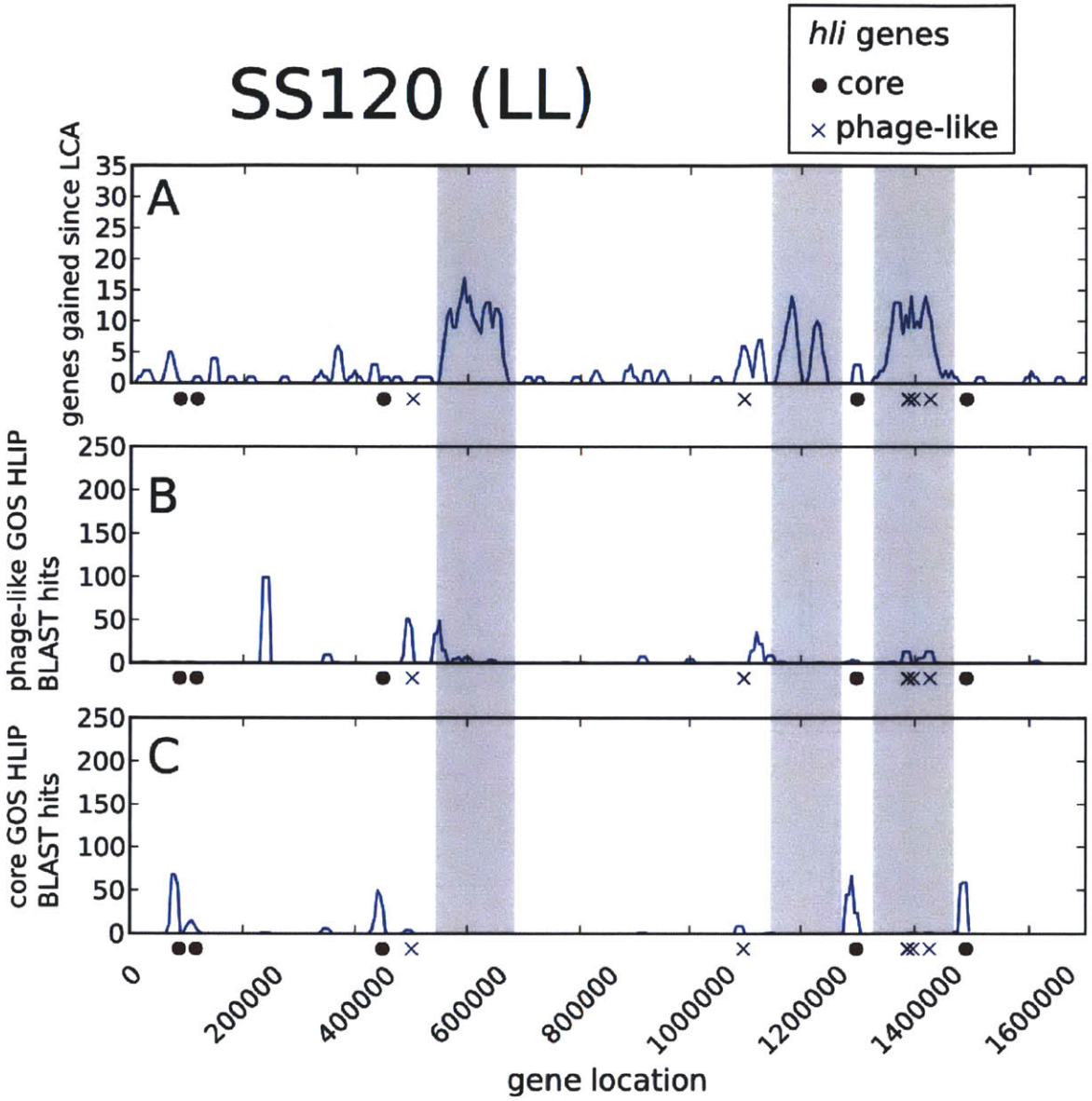


Figure 4-9: Locations of BLAST hits of HLIP-encoding GOS inserts on the low light-adapted *Prochlorococcus* SS120 genome. Layout is the same as in figure 4-3.

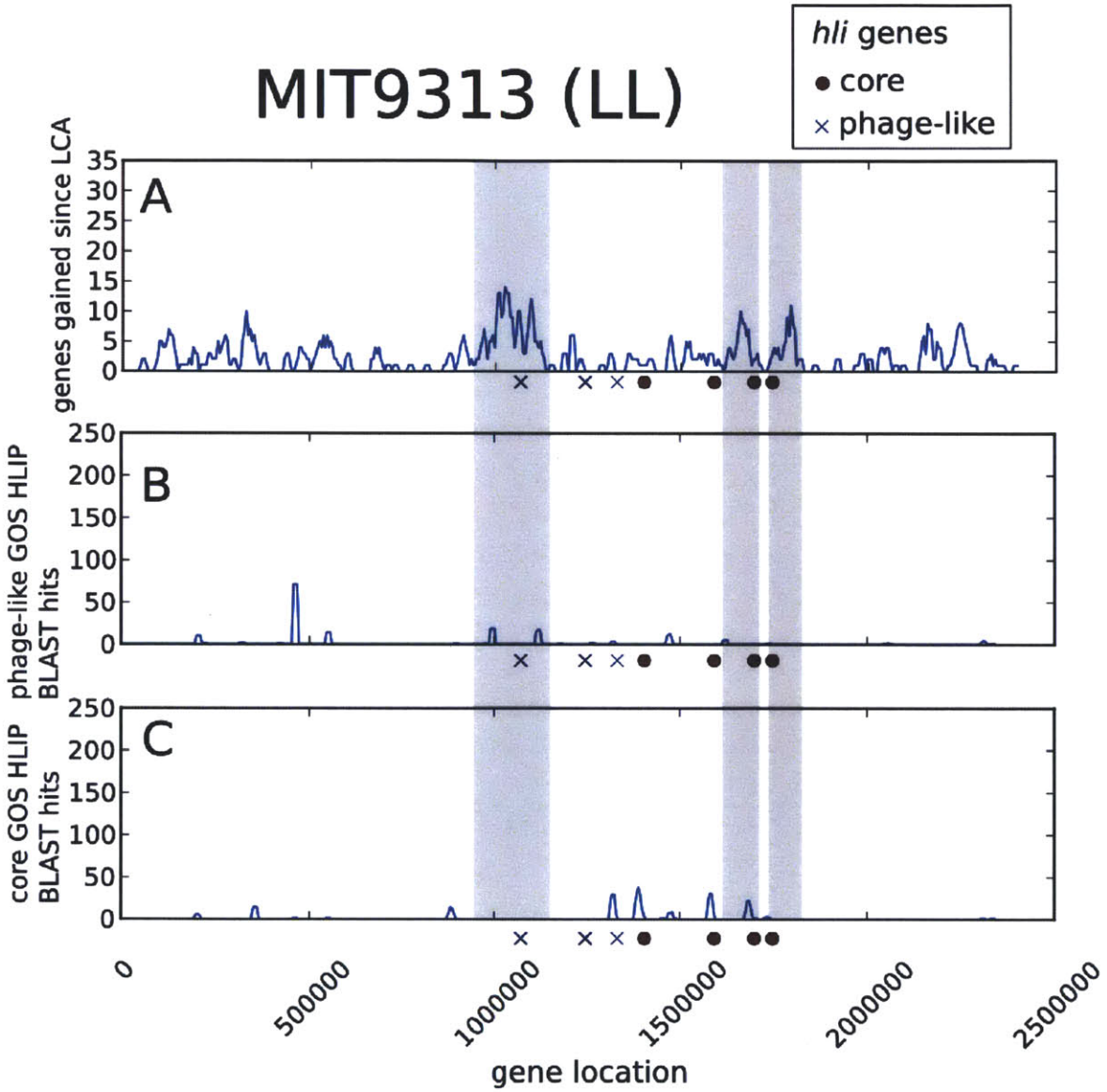


Figure 4-10: Locations of BLAST hits of HLIP-encoding GOS inserts on the low light-adapted *Prochlorococcus* MIT9313 genome. Layout is the same as in figure 4-3.

Chapter 5

Conclusion and future directions

5.1 Conclusion

From its beginning, this thesis was an attempt to bridge the divides between three often-too-distant domains: genomic data, those phenotypes that are assayed in the lab, and environmental data. Chapter 2 was a first step toward that goal: a list of genes in common, and of genes that might define ecotypes, and of genes that define still smaller units of phylogeny. That investigation and Coleman and Chisholm (2007) pointed to a number of possible functional consequences of *Prochlorococcus* genetic diversity, and the subsequent chapters pursued one such lead.

Knowing the genome sequences of NATL isolates, we speculated that their conspicuous number of *hli* genes could have significant functional consequences. While the subsequent chapters focused on HLIPs, and on their putatively niche-defining role as light stress mitigators, it is important to emphasize that they are just one gene family, and eNATL just one ecotype. Research in *Prochlorococcus*, and in environmental microbiology in general, stands to benefit enormously from the sequencing of additional genomes from uncultured single cells. It is very likely that this will produce other candidate genes that define niches we had not considered before. As was the case here, investigations of those genes will benefit if there are straightforward possible laboratory experiments to test genome-based hypotheses – and, of course, only if there are cultured examples of the clade of interest.

The evolution of *hli* genes is curious, and it may be a model for the transfer of other host genes to and from phage. Such transfers have been shown to be important in phage evolution, as the phage uses its host-like proteins to manipulate the cell's core metabolism (Thompson et al., In press). Transfers from phage genomes back to host genomic islands is also important in adaptation

to shortages of nutrients such as phosphate (Coleman et al., 2006). In the case of phosphate limitation, however, those island-borne copies are the only members of their respective families to appear anywhere in the genome (Martiny et al., 2006). In the case of HLIPs, a set of core genes exists in the same genome with their cousins, which were re-acquired from phage, and those separate sets now appear to have separate roles. Did the phage, by maintaining its own HLIP variants, provide a source of diversity that later proved advantagous to the host cell? It will be worth investigating how often this has taken place in other gene families.

It should be emphasized that there are likely to be many, as yet undiscovered gene families that also underlie the phenotypic diversity we have so far observed in *Prochlorococcus*. Likewise, while the tolerance of light shocks in one low light-adapted clade is surprising and worth investigation, it is only one of many possible experiments that might further divide the extant *Prochlorococcus* ecotypes. The true number of functionally distinct *Prochlorococcus* genotypes in the wild is unknown, even if some part of the variation between closely related groups must certainly be genetic drift. Each newly observed, functionally distinct *Prochlorococcus* clade implies it has colonized a niche that would have been closed, or at least a greater challenge, to other *Prochlorococcus*, or indeed other marine microorganisms.

It should also be emphasized that this work focused on one monophyletic group: eNATL. There are an unknown number of other ecologically significant groups that are not monophyletic, such as those cells that possess certain phosphate metabolism genes (Martiny et al., 2009a). This study necessarily focuses on a larger group because the available data (QPCR) is best suited to it, but future studies will benefit from greater metagenomic information and should be able to illuminate more of those scattered genes. Those scattered genes, of course, can be expected to occur in islands. But an ambiguity when we speak of islands will have to be resolved: the data from Chapter 4, Appendix G, and Appendix I suggest that even the major clades are defined by islands; specifically, by island genes that apparently became fixed in a population only once and are now essential to the survival of cells belonging to that clade. That is, different parts of the islands could be most highly variable at different times in *Prochlorococcus* evolution.

5.2 Future directions

5.2.1 Light shock resistance

In chapter 3 I argue that there is at least one significant variation within the LL *Prochlorococcus* clade: the contrast between eNATL and other LL strains during light shocks. However, other variations may exist beyond that. MIT9313 may be somewhat different from SS120 in its initial response to a shock – note the rise in bulk fluorescence in MIT9313ax in the short timecourse, compared to no change at all in SS120. In addition, MIT9313ax may adjust σ, its photosynthetic cross-section, in acclimating to different light levels, a phenomenon not seen in other *Prochlorococcus* (Appendix F). We have not yet seen evidence that this is important to survival in the environment, but as we attempt to find the borders between eMIT9313's niche and that of eSS120, such differences may be worth investigating. Likewise, while chapter 3 treats HL *Prochlorococcus* as a homogeneous group for the purposes of light shock tolerance, that assumption may not hold if probed further. There may be some small difference in light tolerance between MIT9312ax, MED4ax, and MIT9301, for example (Appendix F).

Preliminary data suggests that even when they share the same genes, eNATL and other LL cells may upregulate them under different circumstances (Appendix E). This is an additional layer on the complexity that exists between genomes. Examining differences in gene content is a start to comparing two isolates' genomes, but until those genomes are seen "in motion," such differential regulation will be invisible. Future comparisons of the physiology of any two isolates will necessarily involve such experiments. Fortunately, the availability of custom microarrays will cease to be a limitation with the advent of RNA sequencing (Marioni et al., 2008).

The results in Chapter 3 suggest that it is the early fluorescence quenching in HL and eNATL cells that sets them apart from LL cells under high light stress. HLIPs may be involved, but that must be confirmed and their mechanism described in detail. That line of investigation may continue to benefit from the ongoing research in *Synechocystis*, as models of HLIP function are still being tested. However, the possibility does exist that the split between core and multi-copy HLIPs has functional consequences. In that case, the *Synechocystis* research, which necessarily only studies one of those two *hli* sub-families, may have to be confirmed in *Prochlorococcus*.

5.2.2 The challenge of assigning functions to uncharacterized genes

A disturbingly large fraction of the *Prochlorococcus* pan-genome has no known function. Another significant fraction has only a vaguely defined function. In this work, as in so many other genomics investigations, we benefit from genetic screens and functional characterizations carried out by other groups, often many years ago. We also benefit from a known, easily assayed phenotype in the difference in light shock recoveries among the cultured *Prochlorococcus* isolates. While expanded metagenomics and transcriptomic methods will allow us to probe the space of *Prochlorococcus*, the value of such phenotypes cannot be overestimated. Research in *Prochlorococcus*, and in other environmental genomics models, may come to depend on the use of high-throughput, sequencing-based methods to identify areas that are enriched for particular unknown genes. One could then ask what are the environmental differences that caused one gene or set of genes to be enriched, and how can those environmental differences be replicated in the lab?

The availability of routine RNA sequencing and rapid, sensitive measurements like F_v/F_m could be used to identify differing responses between different strains where gross death/survival assays are less useful. With the amount of sequence data bound to become available, it is arguably more important than ever to try additional stress conditions that reveal the functions of more of the many hypothetical genes.

As additional metagenomic and single-cell sequence data begins to accumulate (Rodrigue et al., 2009), the leaves of the tree that went unexplored in chapter 2 – that is, those poorly understood differences in gene content between closest-cousin sequenced isolates – may become less mysterious.

Appendix A

Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution

Author contributions

D.L., J.D.J, M.L.C. and M.B.S. performed experiments. M.E.F., I.M.A., T.R., G.K., R.S., W.R.H., and G.M.C. analyzed microarray data. D.L. and G.K. carried out RT-PCR verification of the microarray analysis. D.L. and S.W.C. wrote the manuscript.

LETTERS

Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution

Debbie Lindell[1]†, Jacob D. Jaffe[3]†, Maureen L. Coleman[1], Matthias E. Futschik[5], Ilka M. Axmann[5], Trent Rector[4], Gregory Kettler[1], Matthew B. Sullivan[1], Robert Steen[4], Wolfgang R. Hess[6], George M. Church[3] & Sallie W. Chisholm[1,2]

Interactions between bacterial hosts and their viruses (phages) lead to reciprocal genome evolution through a dynamic co-evolutionary process[1–5]. Phage-mediated transfer of host genes—often located in genome islands—has had a major impact on microbial evolution[1,4,6]. Furthermore, phage genomes have clearly been shaped by the acquisition of genes from their hosts[2,3,5]. Here we investigate whole-genome expression of a host and phage, the marine cyanobacterium *Prochlorococcus* MED4 and the T7-like cyanophage P-SSP7, during lytic infection, to gain insight into these co-evolutionary processes. Although most of the phage genome was linearly transcribed over the course of infection, four phage-encoded bacterial metabolism genes formed part of the same expression cluster, even though they are physically separated on the genome. These genes—encoding photosystem II D1 (*psbA*), high-light inducible protein (*hli*), transaldolase (*talC*) and ribonucleotide reductase (*nrd*)—are transcribed together with phage DNA replication genes and seem to make up a functional unit involved in energy and deoxynucleotide production for phage replication in resource-poor oceans. Also unique to this system was the upregulation of numerous genes in the host during infection. These may be host stress response genes and/or genes induced by the phage. Many of these host genes are located in genome islands and have homologues in cyanophage genomes. We hypothesize that phage have evolved to use upregulated host genes, leading to their stable incorporation into phage genomes and their subsequent transfer back to hosts in genome islands. Thus activation of host genes during infection may be directing the co-evolution of gene content in both host and phage genomes.

Prochlorococcus is the dominant photosynthetic organism in vast regions of the world's oceans[7], where T7-like podoviruses are also abundant[8]. Therefore this phage–host system is likely to be of great relevance for bacterial and phage global evolution, for modelling their population dynamics, and for understanding the role of phage in the oceanic carbon cycle.

Phages infecting marine cyanobacteria encode a number of host-like genes including photosynthesis and stress-response genes[5,9–11]. Phage photosynthesis genes are expressed during infection while transcripts of homologous genes in the host decline[12,13], and are hypothesized to facilitate production of carbon and energy through cell photosynthesis for optimal phage production[5,10,12–14]. This physiological interdependence between host and phage is likely to have led to the observed prevalence of photosynthesis genes in cyanophage[10,15], providing a reservoir for genetic exchange, and influencing the co-evolutionary process of both host and phage[14,15]. Although the analysis of single genes has provided insight into this dynamic, a systems approach is essential for a broader understanding of this co-evolutionary process. Here we investigate genome-wide transcriptome dynamics of *Prochlorococcus* MED4 and the T7-like podovirus P-SSP7 over the course of infection—the first such detailed view of infection for any lytic host–phage system.

We first characterized the gross features of the lytic cycle (Fig. 1). Phage genomic DNA (gDNA) began to increase, and host gDNA to decrease, 4 h after infection, and phage progeny were first released into the extracellular medium 8 h post infection (Fig. 1a). Phage

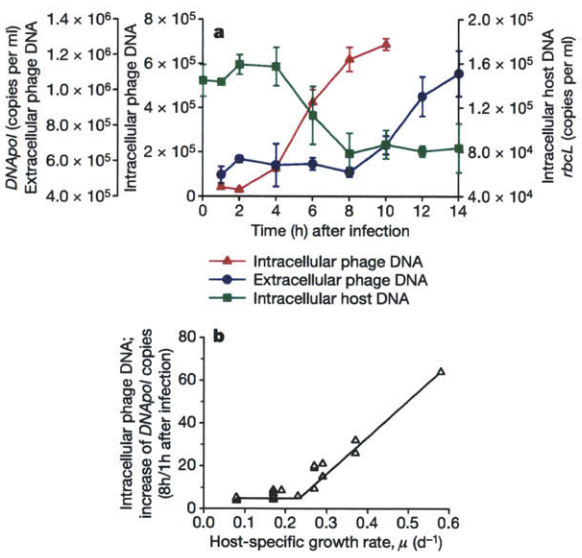


Figure 1 | Infection dynamics of *Prochlorococcus* MED4 by podovirus P-SSP7. a, Timing of phage gDNA replication (intracellular phage DNA) and length of the lytic cycle (extracellular phage DNA) was determined by quantifying the phage DNA polymerase gene (*gene 5*/*DNApol* gene copy number). Host gDNA degradation (intracellular host DNA) was determined by disappearance of the host *rbcL* gene. Average and s.d. of three biological replicates. **b,** Dependence of phage gDNA replication on host growth rate. Phage *DNApol* intracellular copy number was measured 8 h after infection and normalized to that at 1 h after infection as a measure for phage gDNA replication. $n = 24$.

[1]Department of Civil and Environmental Engineering, [2]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [3]Department of Genetics, and [4]BioPolymers Facility, Department of Genetics Harvard Medical School, Boston, Massachusetts 02115, USA. [5]Institute for Theoretical Biology, Humboldt University, Berlin 10115, Germany. [6]Institute of Biology, University of Freiburg, Freiburg 79104, Germany. †Present addresses: Department of Biology, Technion—Israel Institute of Technology, Haifa 32000, Israel (D.L.); The Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02141, USA (J.D.J.).

©2007 Nature Publishing Group

replication is a function of host growth rate (Fig. 1b), which, together with a dependence on photosynthesis[12], suggests an intimate link between phage fitness and host physiology. We wanted to know: what are the dynamics of the phage and host transcriptome during infection, and where do phage-encoded 'bacterial-like' genes fit into this transcriptional program?

The genome content and architecture of the cyanophage P-SSP7 are similar to that of the *Escherichia coli*-infecting T7 podovirus[11]. As in T7 (ref. 16), the P-SSP7 genome was transcribed linearly from the left to the right of the genome map over the course of infection (with important exceptions, see below) (Fig. 2), and three expression clusters were discerned (Fig. 2a, and Supplementary Fig. 1). The first cluster contains a putative *marR* transcriptional regulator gene, a T7 homologue (g0.7) suggesting a role in redirecting transcription from the host towards the phage. The second cluster contains genes involved in DNA metabolism and replication (Fig. 2a, and Supplementary Fig. 1) as well as RNA polymerase (RNAP), which may be involved in RNA transcription and/or DNA replication[16]. The third cluster consists of genes involved in phage particle formation and DNA maturation. Proteins encoded by this latter cluster were detected in the mature phage particle (Fig. 2b, and Supplementary Table 1), further supporting that many are phage structural genes. Thus the three expression clusters in this cyanophage are analogous to T7-coliphage class I, II and III genes in both gene content[11] and the timing of genome expression (Fig. 2). That these fundamental operational properties are conserved across cyanophage and enteric phage, the hosts of which are drastically different with respect to energy source (autotroph versus heterotroph), habitat (nutrient-poor oceanic waters versus the nutrient-rich human gut), and growth rate (generation time of a day versus less than an hour), is remarkable.

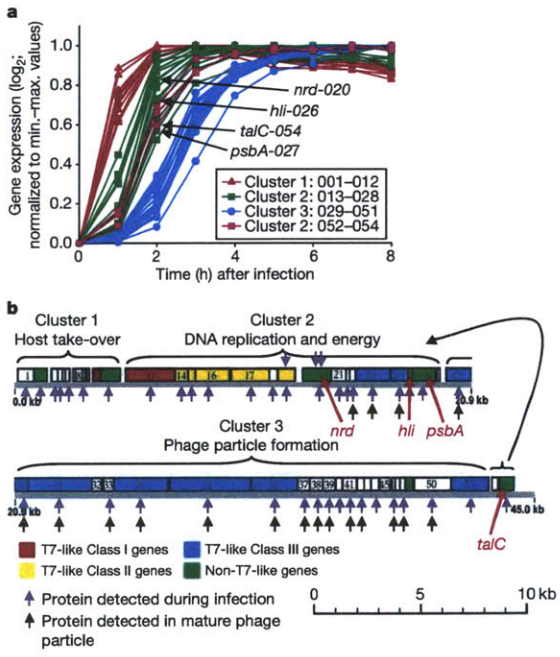


Figure 2 | Temporal expression dynamics of P-SSP7 phage genes during infection of *Prochlorococcus* MED4. a, Transcript levels with time after infection reveal three transcription clusters (see Supplementary Fig. 1). Profiles determined from microarrays, were normalized to minimum–maximum values for each gene. Average of three biological replicates; Supplementary Fig. 6 shows RT–PCR verification of results. **b**, Genome map[11] highlighting the position of *talC* at the end of the genome, even though it is transcribed in cluster 2. Protein detection during infection (purple arrows) and in mature phage particle (brown arrows) showing that 74% of phage genes produced proteins, including three overlapping genes that escaped previous annotation (Supplementary Table 6). Supplementary Table 1 encompasses gene identification and peptide detection.

Despite the similar overall infection strategies of P-SSP7 and T7, transcription cluster 2 in the cyanophage displays novel features in both gene content and regulation and bears signatures of host–phage co-evolution unique to the marine ecosystem. This cluster contains four 'bacterial-like' genes: the ribonucleotide reductase gene *nrd* (ORF 020), the high-light-inducible stress response gene *hli* (ORF 026), the photosystem II gene *psbA* (ORF 027), and the transaldolase gene *talC* (ORF 054). Although *nrd*, *hli* and *psbA* are in the middle of the genome, *talC* is at the end[11] (Fig. 2b). The co-transcription of these four genes, despite their physical separation (Fig. 2, and Supplementary Fig. 2), suggests that they are functionally linked[3].

Clues as to the function of the 'bacterial-like' genes are given by their position in the transcriptional and translational program of the entire host–phage system. First, the proteins encoded by these genes are present during infection but absent from the mature phage particle (Fig. 2b, and Supplementary Table 1), indicating that they function intracellularly. Second, these genes are transcribed together with DNA replication genes, and include ribonucleotide reductase, which converts host ribonucleotides, recycled from degraded RNA (see below), to deoxynucleotides. The photosynthesis genes found in this cluster are thought to be involved in the production of energy[5,10,12–14] and transaldolase may function in the host's pentose phosphate pathway to produce reducing power (NADPH) and/or ribose substrates for nucleotide synthesis[11]. Together, these findings suggest that these genes form a functional unit to produce energy and deoxynucleotide carbon substrates necessary for cyanophage DNA replication in the resource-poor oceans.

The bacterial-like metabolism genes found in P-SSP7 are also commonly found in myoviruses that infect marine cyanobacteria, despite drastic differences in their core genome content[9,11]. In some myoviruses, however, the genes are situated together on the genome[10,11]. Therefore we may be seeing a snapshot of evolution in progress, from spatial separation with cotranscription in P-SSP7, to physically linked genes in other cyanophage genomes.

It is not at all clear how the transcription of this cyanophage genome is regulated, and, in particular, how the last three genes are co-regulated with cluster 2 genes. Although we bioinformatically detected host-like RNAP promoters upstream of each phage expression cluster, and ORF 052, no clear phage-like RNAP promoters were detected[17] (Supplementary Table 2). A transcription initiation site consistent with bacterial-like promoters was experimentally mapped upstream of cluster 2 genes, whereas 5′ ends consistent with RNA processing sites and with weak similarity to T7-like promoters, were found upstream of cluster 1 and cluster 3 genes (Supplementary Fig. 3). However, it remains unclear whether these sequences serve as phage promoters and/or RNA processing sites for transcripts generated by either host or phage RNAP.

Given the reliance of phage replication on host physiology (Fig. 1b), the behaviour of the host transcriptome during infection is of interest. Whereas the transcript levels for approximately 75% of the 1,716 host genes declined during infection (Fig. 3), 41 genes were significantly upregulated. This is distinctly different from other lytic host–phage systems where few, if any, host genes become activated[16,18]. The upregulated genes fall into two groups (Fig. 3, and Supplementary Fig. 4 and Supplementary Table 3). The first was transiently upregulated immediately after infection and consists of high-light-inducible stress response (*hli*), carbon metabolism (*rbcLS*), transcription (*rpoC2*, *rpoD*) and ribosome (*rpl5*, *rpl6*, *rps8*, *rps11*, *rps17*) genes. Transcripts of the second group appeared 2 h after infection and included genes involved in RNA degradation and modification (*rne*, *rnhB*, *dus* and *sun*), protein turnover (*clpS*, and an AAA ATPase family gene), stress responses (*umuD* and *phoH*), and those of unknown function. Two of the latter were

©2007 Nature Publishing Group

transcribed from bacterial-like promoters (Supplementary Fig. 3), suggesting involvement of host RNAP. Upregulated host gene expression may constitute a direct stress response to phage infection, or may have been facilitated by phage factors[16] injected into the cell or expressed from phage expression cluster 1.

Regardless of the mechanism of upregulation, we hypothesize that phages may have evolved to make use of the products of some upregulated host genes as part of the 'arms-race' between host and phage[19]. Certainly, phages are known to exploit host stress-response proteins during infection in other systems[20–22]. The T4 and T7 phages infecting *E. coli*, for example, have evolved to modify host RNase E (involved in RNA degradation) leading to the degradation of host RNA[22]. It is perhaps not coincidental that *rne* (encoding RNase E) is one of the upregulated genes in *Prochlorococcus* during infection. This may have initially served as a host defence mechanism for degrading phage RNA, but could also be exploited by phage to degrade host RNA for use as substrates for phage deoxynucleotide synthesis.

Perhaps the most compelling evidence that upregulated host genes are part of the co-evolutionary process in this system is that 34% of them (more than would occur by chance $P < 0.001$) are found in hypervariable host genome islands (Supplementary Table 3), which are thought to be mobilized by phages[6]. Furthermore, homologues of a number of these host genes are found in phage genomes, including *hli*, *phoH*, and HNH endonuclease and sigma factor genes, as well as RNase H and heat-shock genes[9,11].

Thus there seems to be a connection between genes upregulated during infection, their position in the host genome, and the presence of homologues in phage. Although there are a number of possible explanations for this connection, the most parsimonious evolutionary scenario is as follows: Host stress response genes are upregulated in response to phage infection. Phages that have evolved to use these gene products gain a fitness advantage. Random incorporation[3] of these genes into their own genomes would enable phages to more tightly regulate their expression, conferring a fitness advantage, and leading to preferential retention. This retention would increase the probability of transfer back to the host in genome islands, by lysogeny or unsuccessful infection, and those genes beneficial to the host would remain in the host genome. Analysis of the *hli* gene family provides an interesting illustration in support of this scenario. *hli* genes are upregulated in the host in response to phage infection (Fig. 3, and Supplementary Table 3), are common in *Prochlorococcus*-infecting phage genomes[5,11], and multiple phage-like copies[5] are found in *Prochlorococcus* genome islands[6] (Supplementary Table 4). Furthermore, their differential expression in *Prochlorococcus* in response to various environmental stressors (Supplementary Table 4) and the presence of a binding site for the nitrogen transcriptional regulator NtcA upstream of the nitrogen-regulated *hli10* gene[23], suggests that copies acquired from phage[5] have undergone specialization of function in the host. It remains to be seen whether host fitness has been enhanced by the acquisition of these *hli* genes from cyanophages.

This system-wide analysis of the infection of a cyanobacterium by a phage has led to new insights and hypotheses regarding co-evolutionary interactions between host and phage. These interactions clearly shape the gene content of both host and phage, and probably play a role in shaping the distribution and abundance of cyanobacterial ecotypes in the oceans.

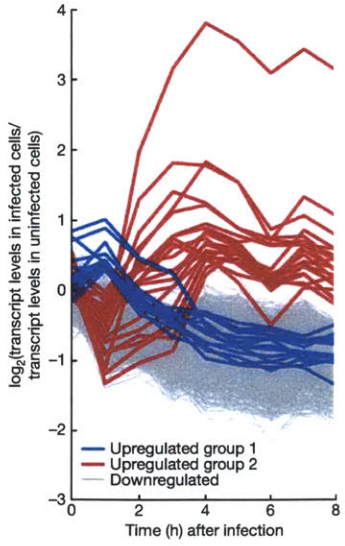


Figure 3 | Transcriptional profiles of *Prochlorococcus* MED4 genes with time after infection by P-SSP7. Transcript levels, determined from microarrays, are presented as $\log_2$-fold change in infected cells relative to uninfected cells over the 8 h latent period of infection. Only genes whose expression levels were significant at a false-discovery rate of $q < 0.05$ are shown. Blue and red indicate significantly upregulated genes in transcription groups 1 and 2, respectively (see Supplementary Table 3). Grey indicates genes significantly downregulated at 8 h after infection. Average of three biological replicates. Supplementary Fig. 7 shows RT–PCR verification of results.

METHODS SUMMARY

Prochlorococcus MED4 was grown at 21 °C under 10–25 μmol photon m^{-1} s^{-1} continuous white light in Pro99 seawater medium with HEPES and sodium bicarbonate. The length of the lytic cycle was determined by quantifying phage DNA in the extracellular medium using a real-time quantitative PCR (qPCR) assay (see Supplementary Fig. 5 for a comparison with standard methods). The timing of phage DNA replication and host DNA degradation were determined intracellularly using qPCR assays for the phage *DNApol* and host *rbcL* genes, respectively. For expression analysis triplicate cultures (10^8 cells ml^{-1}) were infected with the P-SSP7 podovirus (3×10^8 infective phage particles ml^{-1}) and the paired control cultures were amended with filter-sterilized spent medium. Samples were collected by centrifugation, resuspended in storage buffer and snap frozen in liquid nitrogen. RNA was extracted using Ambion's *mir*Vana RNA isolation kit and DNA was removed by DNase I digestion using Ambion's Turbo DNA-*free* kit. Transcriptional analyses were carried out using a custom-made high-density antisense Affymetrix array—MD4-9313. Two micrograms of total RNA were subjected to Affymetrix protocols for *E. coli*. Array analyses were carried out using R and Bioconductor, and array data were normalized and probe set summaries calculated using the robust multi-array average (RMA) procedure[24]. Array results were validated by RT–PCR (Supplementary Figs 6, 7) and the appropriate normalization method was determined by comparing normalized transcription profiles to RT–PCR results (Supplementary Table 5 and Supplementary Figs 8, 9, 10). Promoters were computationally predicted and experimentally assessed using the 5′ RACE technique. For the detection of phage proteins, *Prochlorococcus* cells were harvested 3 and 7 h after infection with phage, and 10^{10} caesium-chloride-purified phage particles were subjected to mass spectrometry proteomic analysis as in ref. 25. See Supplementary Methods for details of all experimental procedures.

Received 5 June; accepted 26 July 2007.

1. Canchaya, C., Proux, C., Fournous, G., Bruttin, A. & Brussow, H. Prophage genomics. *Microbiol. Mol. Biol. Rev.* **67**, 238–276 (2003).
2. Filee, J., Forterre, P. & Laurent, J. The role played by viruses in the evolution of their hosts: a view based on informational protein phylogenies. *Res. Microbiol.* **154**, 237–243 (2003).
3. Hendrix, R. W., Lawrence, J. G., Hatfull, G. F. & Casjens, S. The origins and ongoing evolution of viruses. *Trends Microbiol.* **8**, 504–508 (2000).
4. Hsiao, W. W. L. *et al.* Evidence of a large novel gene pool associated with prokaryotic genome islands. *PLoS Genet.* **1**, e62 (2006).
5. Lindell, D. *et al.* Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc. Natl Acad. Sci. USA* **101**, 11013–11018 (2004).
6. Coleman, M. L. *et al.* Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* **311**, 1768–1770 (2006).
7. Partensky, F., Hess, W. R. & Vaulot, D. *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiol. Mol. Biol. Rev.* **63**, 106–127 (1999).
8. Breitbart, M., Miyake, J. H. & Rohwer, F. Global distribution of nearly identical phage-encoded DNA sequences. *FEMS Microbiol. Lett.* **236**, 249–256 (2004).

©2007 Nature Publishing Group

9. Mann, N. H. *et al.* The genome of S-PM2, a "photosynthetic" T4-type bacteriophage that infects marine *Synechococcus*. *J. Bacteriol.* **187,** 3188–3200 (2005).
10. Millard, A., Clokie, M. R., Shub, D. A. & Mann, N. H. Genetic organization of the *psbAD* region in phages infecting marine *Synechococcus* strains. *Proc. Natl Acad. Sci. USA* **101,** 11007–11012 (2004).
11. Sullivan, M. B., Coleman, M., Weigele, P., Rohwer, F. & Chisholm, S. W. Three *Prochlorococcus* cyanophage genomes: Signature features and ecological interpretations. *PLoS Biol.* **3,** e144 (2005).
12. Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M. & Chisholm, S. W. Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* **438,** 86–89 (2005).
13. Clokie, M. R. J., Shan, J., Bailey, S., Jia, Y. & Krisch, H. M. Transcription of a 'photosynthetic' T4-type phage during infection of a marine cyanobacterium. *Environ. Microbiol.* **8,** 827–835 (2006).
14. Zeidner, G. *et al.* Potential photosynthesis gene recombination between *Prochlorococcus* and *Synechococcus* via viral intermediates. *Environ. Microbiol.* **7,** 1505–1513 (2005).
15. Sullivan, M. B. *et al.* Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biol.* **4,** e234 (2006).
16. Molineux, I. in The Bacteriophages (ed. Calendar, R.) 277–301 (Oxford University Press, New York, 2005).
17. Chen, Z. & Schneider, T. D. Information theory based T7-like promoter models: classification of bacteriophages and differential evolution of promoters and their polymerases. *Nucleic Acids Res.* **33,** 6172–6187 (2005).
18. Miller, E. S. *et al.* Bacteriophage T4 genome. *Microbiol. Mol. Biol. Rev.* **67,** 86–156 (2003).
19. Lenski, R. E. & Levin, B. R. Constraints on the coevolution of bacteria and virulent phage: A model, some experiments, and predictions for natural communities. *Am. Nat.* **124,** 585–602 (1985).
20. Tabor, S., Huber, H. E. & Richardson, C. C. *Escherichia coli* thioredoxin confers processivity on the DNA polymerase activity of the gene 5 protein of bacteriophage T7. *J. Biol. Chem.* **262,** 16212–16223 (1987).
21. Tilly, K., Murialdo, H. & Georgopoulos, C. P. Identification of a second *Escherichia coli groE* gene whose product is necessary for bacteriophage morphogenesis. *Proc. Natl Acad. Sci. USA* **78,** 1629–1633 (1981).
22. Ueno, H. & Yonesaki, T. Phage-induced change in the stability of mRNAs. *Virology* **329,** 134–141 (2004).
23. Tolonen, A. C. *et al.* Global gene expression of *Prochlorococcus* ecotypes in response to changes in nitrogen availability. *Mol. Systems Biol.* **2,** 53 (2006).
24. Irizarry, R. A. *et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31,** e15 (2003).
25. Jaffe, J. D. *et al.* The complete genome and proteome of *Mycoplasma mobile*. *Genome Res.* **14,** 1447–1461 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Steglich, S. Bhattacharya, H. Keller, L. Thompson, P. Weigele, S. Choe, D. Endy, S. Kosuri, M. Shmoish and J. Aach for discussions, and J. Waldbauer and M. Osburne for comments on the manuscript, and the MIT Center for Environmental Health Sciences. This work was funded by the DOE Genomes to Life System Biology Center Grant (G.M.C. and S.W.C.), the Gordon and Betty Moore Foundation's Marine Microbiology Program (S.W.C.), and the National Science Foundation (S.W.C).

Author Information The microarray data have been deposited in the GEO database under the accession number GSE8382. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.W.C. (chisholm@mit.edu).

©2007 Nature Publishing Group

SUPPLEMENTARY INFORMATION

Supplemental Methods

Culture Growth and Experimental Design

Prochlorococcus MED4 was grown in the Pro99 seawater based medium amended with 10 mM HEPES (pH7.5) and 12 mM sodium bicarbonate at 21 °C under continuous white light at 10-25 μmol photon·m^{-1}·s^{-1} as described in Lindell et al.[12]. For experiments in which host gDNA was quantified and expression analyses carried out, the ratio of infective phage to host (the MOI) was 3.0 to maximize levels of infection. Phage at $3x10^8$ infective particles·ml^{-1} were added to 10^8 cells·ml^{-1} and samples were collected each hour during the course of infection. Prior to phage addition for the expression experiment, cells were concentrated to 10^8 cells·ml^{-1} by centrifugation. Experiments were carried out with triplicate independent cultures in paired experiments and control treatments were amended with filter-sterilized spent medium. For experiments in which the length of the lytic cycle and the timing of phage gDNA replication were determined, the ratio of infective phage to host was 0.1. Phage at 10^7 infective particles·ml^{-1} were added to 10^8 cells·ml^{-1} and allowed to adsorb for 1 h. The phage-cell mix was then diluted 100 fold and samples collected at different times after phage addition.

Quantification of phage particles and phage and host genomic DNA

Extracellular Phage Quantification

Phage particles from the extracellular medium were quantified using a quantitative PCR (qPCR) method. Samples were filtered over 0.2 μm sterile syringe filters (Tuffryn HT), and the filtrate, containing phage particles, was collected. To prevent PCR inhibition by the seawater based growth medium, the filtrate was diluted 20-100 fold in 10 mM Tris pH8, 10 μl of which was used in triplicate qPCR assays with the P-SSP7 specific DNA polymerase primers (see Suppl. Table 7 for primer sequences). Quantification was achieved using a standard curve of phage particles in 10 mM Tris pH8 that had been enumerated by epifluorescence microscopy after SYBR staining (see below).

This qPCR method was compared to standard methodology for determining phage numbers in the extracellular medium (Suppl. Fig. 5). The number of infective phages was determined by the Most Probable Number (MPN) assay[26]. Briefly, phage samples were serially diluted and added to exponentially growing MED4 cells in 96-well plates. The clearing of wells, as compared to control wells, was monitored using a Synergy HT Biotek fluorescence plate reader. The number of cleared wells at the appropriate phage dilutions was used to calculate the most probable number of infective phage in the undiluted sample. Total phage particles were enumerated using epifluoresence microscopy after DNA staining of the phage[27]. Briefly phage samples were filtered onto a 0.02 μm Anodisc (Whatman) filter with a vacuum of 7 in of Hg and allowed to dry. The filter was then stained with SYBR Green I (Molecular Probes), and the sample enumerated by epifluorescence microscopy after addition of anti-fade solution containing 0.1% p-phenylenediamine.

Intracellular phage and Prochlorococcus DNA quantification

Prochlorococcus cells were collected onto 0.2 μm pore-sized polycarbonate filters (Osmonics) by filtration at 8-10 in of Hg. The filters were washed 3 times with sterilized

seawater to reduce the presence of extracellular phage, once with 3 ml preservation solution (10 mM Tris, 100 mM EDTA, 0.5 M NaCl; pH8) and then frozen at -80 °C. A heat lysis method was used to extract DNA from *Prochlorococcus* cells[28]. Briefly, the polycarbonate filter with *Prochlorococcus* cells was immersed in 650 µl of 10 mM Tris pH8, and agitated in a mini-bead beater for 2 min at 5000 rpm without beads. Five hundred µl of the sample was removed from the shards of filter and heated at 95 °C for 15 min. Ten µl was used in triplicate qPCR reactions. Phage DNA was amplified with P-SSP7 specific DNA polymerase primers, and *Prochlorococcus* DNA with *rbcL* primers (see Suppl. Table 7 for primer sequences).

Quantitative PCR protocol

Triplicate real-time PCR assays were carried out for each sample using Qiagen's QuantiTect SYBR Green PCR kit, primers at 0.3-1.0 µM and 10 µl samples (in 10 mM Tris pH8) in 25 µl volume reactions run on an DNA Engine Opticon (MJ Research). After 15 min denaturation at 95 °C, 40 amplification cycles were carried out as follows: Denaturation (95 °C for 15 sec); annealing (56 or 58 °C for 30 sec); elongation (72 °C for 30 sec); and fluorescence plate read (for quantification of SYBR green incorporation into double stranded DNA), and were followed by 5 min at 72 °C and melt curve analysis (read every degree from 50-90°C). Quantification of template was determined from standard curves produced with dilution series of P-SSP7 phage particles or *Prochlorococcus* MED4 genomic DNA. Melt curve analysis was used to verify that a single product was amplified.

RNA extraction

Samples were collected by centrifugation (12,400 Xg for 15 min at 20 °C), resuspended in storage buffer (200 mM sucrose, 10 mM sodium acetate pH5.2, 5 mM EDTA), snap frozen in liquid nitrogen and stored at -80 °C. Prior to RNA extraction the storage buffer was removed after spinning the cells for 2 min at 20,000 Xg at 20 °C. RNA was extracted using Ambion's *mir*Vana RNA isolation kit. DNA was removed by DNase I digestion using the Turbo DNA-*free* kit (Ambion). For microarray analysis 8 µg of the nucleic acid extract was digested with 6 U of Turbo DNase during a 60 min incubation at 37 °C followed by DNase I inactivation with inactivation slurry. The RNA was purified and concentrated by sodium acetate/ethanol precipitation. DNA removal was verified by gel electrophoresis. For RT-PCR analysis DNA was removed from 0.1-0.5 µg of the nucleic acid extract using the above procedure but without the precipitation step. DNA removal was verified by running no RT controls followed by qPCR for each sample (see RT-PCR validation of array results).

Array experimentation

Transcriptional analysis was carried out using a custom-made high density antisense Affymetrix array – MD4-9313. Synthesis of complementary DNA (cDNA), labeling, hybridization, staining and scanning was carried out according to Affymetrix protocols for *E.coli* (http://www.affymetrix.com/support/technical/manual/expression_manual.affx) with minor changes. Total RNA (2 µg) was denatured at 70 °C and annealed to random hexamer primers (25 ng/µl) at 25 °C for 10 min. The RNA was reverse transcribed to produce cDNA with Superscript II (25 U/µl – Invitrogen Life Technologies) and 0.5 mM

dNTPs in the presence of 1 U/μl RNase Out RNase Inhibitor (Invitrogen). The mix was incubated at 25 °C for 10 min followed by 60 min incubations at 37 °C and 42 °C respectively. Superscript II was inactivated with a 10 min incubation at 70 °C. Sodium hydroxide (0.25 N) was used to remove RNA during a 30 min incubation at 65 °C, followed by neutralization with HCl. The cDNA was purified with MinElute PCR purification columns (Qiagen). Fragments of cDNA, 50-200 nt long, were produced from a 10 min incubation at 37 °C with DNase I (0.6 U per μg cDNA), followed by heat inactivation of the DNase I enzyme (10 min at 98 °C). The cDNA fragments were end-labeled with biotin using the BioArray Terminal Labeling Kit (Enzo) during a 60 min incubation at 37 °C. The reaction was stopped by freezing at –20 °C. The quality of biotin end-labeling was verified by gel-shift assays with NeutrAvidin (Pierce Chemicals) on 1% TBE agarose gels.

The cDNA was hybridized to the MD4-9313 custom Affymetrix array (see below for array description) in aqueous hybridization solution (100 mM MES, 1 M NaCl, 20 mM EDTA, 0.01% Tween-20, 0.1 mg/mL Herring Sperm DNA, 0.5 mg/mL BSA, 7.8 % DMSO and 3 nM prelabeled Affymetrix hybridization B2 oligo control probe mix) during a 16 h incubation at 45 °C in a GeneChip Hybridization Oven 320 rotating at 60 rpm. Washes and stains were carried out on a GeneChip Fluidics Station 450 (Affymetrix) following the ProkGE_WS2v3 Affymetrix protocol. Briefly, following two stringency washes the array was sequentially incubated with 10 μg/mL streptavidin (Pierce Chemical), 5 μg/mL biotinylated anti-streptavidin goat antibody (Vector Laboratories) and 0.1 mg/mL goat IgG (Sigma) and 10 μg/mL streptavidin-phycoerythrin conjugate (Mol. Probes) each for 10 min at 25 °C. After a final wash the arrays were scanned with the GeneChip Scanner (Affymetrix) at a 2.5 μm resolution with excitation set for 570 nm.

Array Design

The MD4-9313 (MD4-9313a520062) array is a custom-made high density antisense Affymetrix array. This array detects labeled cDNA that is antisense to the original RNA. It contains probes for the genomes of two cyanobacterial strains, *Prochlorococcus* MED4 and *Prochlorococcus* MIT9313[29], as well as two dsDNA phages that infect *Prochlorococcus* MED4 – the podovirus P-SSP7 and the myovirus P-SSM4[11]. For the *Prochlorococcus* genomes, the array contains probe sets to detect all predicted open reading frames with probe pairs approximately every 80 bases. For short open reading frames (ORF), the length of the gap was reduced to ensure a minimum of 11 probe pairs per ORF where possible. Probes were designed for intergenic regions longer than 35 bases and were spaced every ca 45 bases on both strands. For short intergenic regions the gap between probe pairs was reduced to ensure a minimum of 4 probe pairs where possible. For some short sequences, where insufficient high performance probes were designed using this approach, the best probes possible were designed with no regard for their spacing along the genome feature. For the phage isolates, probe pairs were designed across the genomes for both strands at an approximate interval of 90 bases. Probes are 25 bases long and each probe pair consists of a perfect match probe (identical to the sequence) and a mismatch probe (containing a single base change at the center of the probe).

Array Data Analyses

Normalization and Statistical Analyses

Data analyses were carried out in the statistical language R using several Bioconductor packages[30]. The array data were normalized and probe set summaries calculated from perfect match probe intensities in Affymetrix CEL files using quantile robust multi-array average (RMA) analysis[24] as implemented in the Bioconductor package *affy*[31]. See below for determination of appropriate normalization method for this experiment. Statistical significance of differentially expressed genes between infected and control cells at each time point was determined using the Bayesian t-test function implemented in the GoldenSpike[32] package (originally derived from the *hdarray* package) with the confidence level set to 9[32, 33]. The results are comparable to those obtained when RMAExpress Version 0.3 beta 1[24], Cyber-T[34] and Q-value[35] were used as stand-alone programs (data not shown). Control arrays at 4 and 8 h after infection gave the same expression profiles (data not shown) and were used for comparison with infected cells at 5, 6, 7 h after infection.

Clustering Analyses

Hierarchical clustering of phage genes was carried out using Pearson correlation and average linkage in the *stats* package in R. Input data was the average logged expression values of 3 biological replicates, standardized so that mean expression values for each gene equal zero and standard deviation equals one. The dendogram, visualized with Java TreeView[36], suggests the presence of several distinct expression clusters (Suppl. Fig. 1a). To determine the number of reliable clusters in the data, a resampling approach was applied[37] using the Bioconductor package *clusterStab*[38] whereby randomly selected sub-sets of genes are repeatedly clustered and the extent of similarity between the resulting clusters are examined. Reliable (or stable) clusters are those which repeatedly occur for the random sub-sets of genes. The similarity between clusters of different repeats was measured by the Jaccard coefficient ranging from zero (no similarity) to one (identical clustering). This resampling strategy was used for a range of number of clusters (k=2 to k=5) and the resulting distribution of Jaccard coefficients compared. If an adequate number of clusters is chosen, the distribution of coefficients will show an enrichment of values equal or close to one. A comparison of the histograms for the Jaccard coefficients strongly indicated the existence of three stable expression clusters for P-SSP7 genes (Suppl. Fig. 1b). Their temporal profiles are shown in Suppl. Fig. 1c. While this normalization methodology is the most appropriate for cluster analysis, we show temporal phage gene expression profiles in Fig. 2a using minimum-maximum normalized data as these more appropriately describe the dynamics of phage genome expression from a biological perspective.

The same strategy was used to determine the number of stable clusters for upregulated MED4 genes. Here the histograms indicate that two stable expression clusters exist (Suppl. Fig. 4).

Significance of co-expression of 'bacterial-like' phage genes

Clustering analysis indicated that the 'bacterial-like' phage genes *nrd, hli, psbA* and *talC* are temporally coexpressed (Suppl. Fig. 1, Fig. 2). To stringently assess the validity of

this co-expression, two approaches were used. First, we assessed the reliability that the four genes are assigned to the same expression cluster (namely cluster 2) by a bootstrap approach using the Bioconductor package *hopach*[39] whereby genes are assigned to a particular cluster when only partial time series are used. Reliable cluster assignments should not depend on single data points and should therefore be found using only partial data. Thus, reliability can be examined by repeated bootstrap sampling and re-clustering of genes with subsequent calculation of cluster memberships. Cluster membership is defined here as the percentage of bootstrap samples that were assigned to the same original cluster. A cluster membership close to one indicates reliable assignment of a gene to the cluster. Applying this bootstrap approach, we detected high membership values for the phage genes *nrd, hli, psbA* and *talC* (1.000, 0.998, 0.9984 and 0.9999) for cluster 2 (Suppl. Fig. 2a). This strongly indicates that the 4 'bacterial-like' phage genes are co-transcribed together within cluster 2 despite their spatial separation on the genome.

In addition, a regression approach was applied to determine the degree of certainty of co-expression of the four 'bacterial-like' genes (Suppl. Fig. 2b, 2c). In this analysis we wished to estimate the time points at which expression of each P-SSP7 gene changed from being non-expressed to expressed – termed here the switch time t*. If genes are co-regulated, we expect them to have the same time t* within acceptable confidence intervals. Here we defined the switch time t* as the time at which expression values reach half maximum values. We used the following procedure to estimate t*: After averaging expression values for the 3 biological replicates, values for each gene across the time series were normalized to a minimum value of zero and maximum value of 1. All P-SSP7 genes displayed sigmoidal expression patterns. To improve the fitting by sigmoidal curves, expression values across the time series were truncated to values ranging from 0.01 to 0.95 .The truncated expression values y were fitted to the sigmoidal function; $y(t) = \exp(a.t+b)/(\exp(a.t+b) + 1)$, where a and b are fitting parameters and t is the time after infection. To allow fitting of the data by linear regression (which simplifies the calculation of confidence intervals), the data were transformed such that $y' = \log(y/(1-y))$. Subsequently, linear regression given by $y' = a.t + b$, was performed and confidence intervals were calculated for each gene. Seeing as $y=0.5$ (half maximal expression) corresponds to $y'=0$, we can use the confidence intervals for y' to assess whether the induction of genes occurred at the same time t*. The confidence intervals for *nrd* (020), *hli* (026), *psbA* (027) and *talC* (054) all overlap indicating that they are turned on simultaneously, together with the remaining genes in cluster 2 (Suppl. Fig. 2b). An example of the fitting procedure is illustrated in Suppl. Fig. 2c for the *nrd* gene.

Determination of Appropriate Normalization Method

Analysis of microarray data after implementation of various normalization methods showed differences in putative expression patterns, in particular for down-regulated genes (Suppl. Fig. 8). Therefore to ascertain which normalization method should be used for this dataset, normalized expression patterns for select genes were compared to those determined empirically with RT-PCR (see below for RT-PCR methodology). Normalization procedures tested were: RMA with quantile normalization at the probe level; RMA with normalization based on positive hybridization control spikes (AFFX-Bio* and AFFX-Cre*); Goldenspike which computes an expression summary based on 8

different normalization methods[32]; and Goldenspike without the second loess normalization at the summary level – as the assumption for this summary level normalization, that the majority of genes is not differentially expressed, may not hold for this experiment.

Comparisons were carried out on representative genes with the following expression patterns: (a) 1 unchanged, internal control gene; (b) 4 down-regulated genes and (c) 7 up-regulated genes. See Suppl. Table 5 for a list of the genes tested. (Note that PMM0550 and PMM1629 are considered both up- and down-regulated based on RMA quantile normalization.) See the "RT-PCR validation of array results" section below for more details of the genes chosen for RT-PCR validation. We compared the performance of the different normalization schemes to detect differential expression as validated by RT-PCR. These analyses show that RMA and the two versions of Goldenspike performed similarly for up-regulated expression, whereas differential expression patterns for down-regulated genes were best represented by RMA with quantile normalization (Suppl. Table 5, and see Suppl. Figs. 8, 9).

Initially, the superior performance of RMA with quantile normalization was somewhat surprising, seeing as it assumes similar overall distribution of probe intensities in different arrays, and we observed downregulation of a large number of genes. However it is important to note that a considerable percentage (25%) of the host MED4 genes were not significantly down-regulated at 8 h after infection. More importantly, however, is that the array used in this study includes a large number of probe sets other than for the host genes. It contains probe sets for intergenic regions, for the P-SSP7 phage genome and for an additional *Prochlorococcus* and phage strain. In fact, most probes on the microarray are not assigned to the organisms examined in our study. Therefore, most expression signals on the array are not expected to change and the underlying assumption of quantile normalization may hold. To assess this issue further, we compared the frequency of probe intensities for the whole array to the subset of intensities for MED4 genes after normalization. The density plots show that the signal distributions for the subset of MED4 probe sets are distinct for different arrays even though the overall distributions are similar for all arrays due to quantile normalization (Suppl. Fig. 10). Thus, quantile normalization can still be applied to our study as it did not erase differences in expression for the host genes.

RT-PCR validation of array results

Total RNA (2-5 ng) was reverse transcribed with Superscript II (Invitrogen) using 2 pmol gene-specific reverse primers in 20 µl reactions following the manufacturer's instructions. The resultant cDNA was diluted with 80 µl 10 mM Tris pH8, and 10 µl was used in each of 3 triplicate quantitative PCR reactions using Quantitect Sybr Green 2x kit (Qiagen) and 0.5-1.0 µM primers (see Suppl. Table 7 for primer sequences) in 25 µl reactions, such that cDNA resulting from 0.2-0.5 ng total RNA was used in each qPCR reaction (see above for quantitative PCR protocol). Results were further normalized to *rnpB* transcript levels which served as an internal control. No RT controls, carried out under identical conditions but without the reverse transcriptase enzyme, indicated that gDNA contamination was less than 1 % of the RT-PCR signal in all samples. Standard

curves were carried out with MED4 gDNA or P-SSP7 phage particles. Expression levels from infected cells at different times after infection was compared to that for control cells. Significance of differential expression was determined from two-tailed t-tests.

Phage genes chosen for RT-PCR validation included the first gene of each expression cluster as well as the last gene in the genome (*talC*) as a representative of the 3 genes transcribed out of order on the genome. RT-PCR validation of host genes included downregulated and upregulated genes from both up-regulated expression clusters and were chosen to span low, medium and high array signal intensities as well as to include genes of potential biological interest where possible.

Promoter analysis

Bioinformatic analysis of phage P-SSP7 transcriptional signals

The computational prediction of bacterial promoters was based on a position specific weight matrix established for the -10 box of *Prochlorococcus* MED4 promoters[40]. Bacterial terminators were found by using the TransTerm algorithm[41], which detects rho-independent transcription terminators by searching for stem-loop-structures (inverted repeats) followed by a row of T's in the genome. Putative recognition sites for the phage RNA polymerase were searched *in silico* with a consensus sequence for T7 RNA polymerase allowing substitutions at positions, which are not common among all 47 natural phage promoters[42].

Experimental detection of 5' transcript ends

The 5' ends of mRNA transcripts from P-SSP7 and MED4 were mapped using the 5' Rapid Amplification of cDNA ends (RACE) technique, described previously by Bensing et al.[43] and modified for *Prochlorococcus* by Vogel et al.[40]. Briefly, 0.7-1.5 μg total RNA was used to cleave the 5' triphosphate, found in primary transcripts, with tobacco acid pyrophosphatase (TAP) (Epicentre, Madison, Wisconsin USA). The resulting 5' monophosphate was subsequently ligated, using T4 RNA ligase (Epicentre, Madison, Wisconsin USA), to the 3' hydroxyl group of an RNA oligonucleotide (5' adaptor: GAU AUG CGC GAA UUC CUG UAG AAC GAA CAC UAG AAG AAA). A gene-specific DNA primer (see Suppl. Table 8 for gene specific primer sequences) was used for reverse transcription followed by PCR amplification with a nested gene-specific primer and the 5' adaptor primer (ATA TGC GCG AAT TCC TGT AGA ACG AAC ACT AG). The amplification products were cloned and sequenced, and the first nucleotide downstream of the 5' adaptor RNA was assigned as the 5' end. This method enables the differentiation between transcription initiation sites of primary transcripts and RNA processed sites. For primary transcripts (carrying a 5' triphosphate) the TAP treated samples (TAP+) yield a specific or strongly enhanced amplification product relative to untreated samples (TAP-), whereas amplification products of equal intensity found for both TAP treated and untreated RNA samples are indicative of processed 5' ends that already carried a monophosphate at the 5' end.

Protein Analyses

Protein Extraction and Digestion to Peptides

Cells were collected by centrifugation and stored as described for RNA work, prior to lysis in 3 M urea, 0.05% SDS, and 50 mM Tris-HCl pH 8.0. Protein levels were quantified using the bicinchoninic acid method (Pierce). Samples were digested with sequencing grade trypsin (Promega) (protein:trypsin = 137.5:1) overnight at 37 °C, reduced with 10 mM DTT, alkylated with 50 mM iodoacetamide, and acidified to < pH 3.0 with HCl.

Peptide Fractionation and identification by Ion Trap Mass Spectrometry (MS)

Two phage infection time points (3 h and 7 h post infection) were subjected to comprehensive MS/MS sequencing experiments. For this purpose, each sample was adjusted to 25% acetonitrile and centrifuged to remove particulates. The entire sample was subjected to two-dimensional chromatographic fractionation (strong cation exchange followed by reversed phase) as in Jaffe et al. [44]. The eluate of the nano-flow reversed phase column was coupled directly to a LTQ linear ion trap mass spectrometer (ThermoElectron, Waltham, MA) where the top 10 most abundant MS ions were sampled for MS/MS sequencing in each scan cycle. Dynamic exclusion was employed to increase depth of coverage. In all, 60 Strong Cation Exchange (SCX) fractions were analyzed for each sample. The accumulated spectra were analyzed with SEQUEST[45], searching against a database of all predicted proteins from *Prochlorococcus* MED4 as well as the entire genomic sequence of cyanophage P-SSP7 prepared for proteogenomic mapping as in Jaffe et al.[44]. Criteria for valid spectra assignments and creation of proteogenomic maps for the phage were as in Jaffe et al.[44]. All validated peptides were considered to be potentially 'present' in subsequent analyses. It should be noted that detection of peptides is dependent on its ionization properties and on it being of suitable length, therefore the inability to detect a particular peptide using this methodology is not a definitive indication of the lack of its presence.

Phage Particle Purification for protein analysis

Prochlorococcus MED4 was infected with P-SSP7 and harvested once the culture had cleared. The cell lysate was centrifuged at 12,000 Xg for 30 min to remove unlysed cells and cellular debris. The supernatant was incubated for 30 min at 25 °C with 1 μg DNase I (Sigma) to degrade host gDNA from lysed cells. The salt concentration of the solution was brought up to 2 M with NaCl and incubated at 25 °C for an additional 30 min and then spun at 15,000 Xg for 30 min and the pellet discarded. Triton X-100 (0.1 % v/v final concentration) and PEG 8000 (10 % w/v final concentration) was added to the supernatant and stirred gently until fully dissolved and incubated overnight at 4 °C. Phages were collected by centrifugation at 12,400 Xg for 30 min at 4 °C and resuspended in Pro99 medium. Phage particles were purified on a cesium chloride step gradient (rho=1.4/1.6) prepared in 0.2 μm filtered seawater amended with 50 mM $MgCl_2$, 50 mM Tris pH8 and 0.1 % Triton X-100, and spun at 150,000 Xg for 2 hours. Purified phage particles were dialyzed in a step-wise fashion against 1 l of 1M NaCl, 50 mM $MgCl_2$, 50 mM Tris pH8 for 1 hour and twice for an hour each against 1 l of 100 mM NaCl, 50 mM $MgCl_2$, 50 mM Tris pH8.

Determination of proteins in purified phage particles

The equivalent of 10^{10} purified phage particles was subjected to proteomic analysis. The sample was digested for 18 hours at 37 °C with 0.2 μg of trypsin in a buffer consisting of 3M urea, 25 mM Tris pH 8.0, 25 mM $MgCl_2$, and 50 mM NaCl. The sample was reduced and alkylated as above, except that 5 mM DTT and 12.6 mM iodoacetamide were used. The sample was desalted using an Oasis HLB solid phase extraction column (Waters, 10 mg resin) according to the manufacturer's directions, reduced to dryness by vacuum centrifugation, and resuspended in 10 μl of 5% acetonitrile/5% formic acid. The sample was analyzed with 3 x 3μl injection to LCMS as above using a top 10 MS/MS method on an LTQ-FT mass spectrometer. Spectra were extracted and searched using SpectrumMill (Agilent, Palo Alto, CA) against the same hybrid database of phage and host proteins described above. Standard SpectrumMill autovalidation parameters were used to select confidently identified proteins and peptides.

Supplemental References

26. Taylor, J. The estimation of numbers of bacteria by tenfold dilution series. Journal of Applied Bacteriology 25, 54-61 (1962).
27. Noble, R. T. & Fuhrman, J. A. Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. Aq. Microb. Ecol. 14, 113-118 (1998).
28. Zinser, E. R. et al. *Prochlorococcus* ecotype abundances in the North Atlantic Ocean as revealed by an improved quantitative PCR method. Applied and Environmental Microbiology 72, 723-32 (2006).
29. Rocap, G. et al. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature 424, 1042-1047 (2003).
30. Gentleman, R. C. et al. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology 5, R80 (2004).
31. Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. Affy - analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20, 307-315 (2004).
32. Choe, S. E., Boutros, M., Michelson, A. M., Church, G. M. & Halfon, M. S. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. Genome Biology 6, R16 (2005).
33. Baldi, P. & Long, A. D. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. Bioinformatics 17, 509-519 (2001).
34. Long, A. D. et al. Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Journal of Biological Chemistry 276, 19937-19944 (2001).
35. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences U S A 100, 9439-9445 (2003).
36. Saldanha, A. J. Java Treeview - extensible visualization of microarray data. Bioinformatics 20, 3246-3248 (2004).
37. Ben-Hur, A., Elisseeff, A. & Guyon, I. A stability based method for discovering structure in clustered data. Pacific Symposium on Biocomputing (2002).
38. Smolkin, M. & Ghosh, D. Cluster stability scores for microarray data in cancer studies. BMC Bioinformatics 4, 36-42 (2003).

39. Pollard, K. S. & van der Laan, M. J. Cluster Analysis of Genomic Data. *In:* Bioinformatics and Computational Biology Solutions Using R and Bioconductor; R. Gentleman, V. Carey, W. Huber, R. Irizarry, S. Dudoit (eds.) Springer, 209-229 (2005).
40. Vogel, J., Axmann, I. M., Herzel, H. & Hess, W. R. Experimental and computational analysis of transcriptional start sites in the cyanobacterium *Prochlorococcus* MED4. Nucleic Acids Research 31, 2890-9 (2003).
41. Ermolaeva, M. D., Khalak, H. G., White, O., Smith, H. O. & Salzberg, S. L. Prediction of transcription terminators in bacterial genomes. J. Mol. Biol. 301, 27-33 (2000).
42. Imburgio, D., Rong, M., Ma, K. & McAllister, W. T. Studies of promoter recognition and start site selection by T7 RNA polymerase using a comprehensive collection of promoter variants. Biochemistry 39, 10419-10430 (2000).
43. Bensing, B. A., Meyer, B. J. & Dunny, G. M. Sensitive detection of bacterial transcription initiation sites and differentiation from RNA processing sites in the pheromone-induced plasmid transfer system of *Enterococcus faecalis*. Proceedings of the National Academy of Sciences U S A 93, 7794-7799 (1996).
44. Jaffe, J. D., Berg, H. C. & Church, G. M. Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics 45, 59-77 (2004).
45. Eng, J. K., McCormack, A. L. & Yates, J. R. r. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. Journal of the American Society for Mass Spectromotry (1994).
46. Steglich, C., Futschik, M., Rector, T., Steen, R. & Chisholm, S. W. Genome-wide analysis of light sensing in *Prochlorococcus*. Journal of Bacteriology 188, 7796-7806 (2006).
47. Dunn, J. J. & Studier, F. W. Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. J. Mol. Biol. 166, 477-535 (1983).

Supplementary Table 1: Detection of phage proteins inside the host cell during infection (infection) or in the purified phage particle (virion). Number of unique peptides detected: 1 = +; 2-4 = ++; 5-10 = +++; more than 10 = ++++. Transcription cluster designations as per Fig.2 and Suppl. Fig. 1.

ORF ID	Gene Name – Product	Infection	Virion	Transcription cluster
PSSP7_001	unknown	++		Cluster 1
PSSP7_002	unknown	++		
PSSP7_003	unknown	++		
PSSP7_004	unknown	++		
PSSP7_005	unknown			
PSSP7_006	unknown	+++		
PSSP7_007	unknown			
PSSP7_008	unknown			
PSSP7_009	unknown	+		
PSSP7_010	unknown			
PSSP7_011	*gene 0.7* – MarR family of transcriptional regulators			
PSSP7_012	*int* – integrase	+		
PSSP7_013	*gene 1* – RNA polymerase	++		Cluster 2
PSSP7_014	*gene 2.5* – ssDNA binding protein	+++		
PSSP7_015	*gene 3* – endonuclease	+		
PSSP7_016	*gene 4* – primase/helicase	++++		
PSSP7_017	*gene 5* – DNA polymerase	+++		
PSSP7_018	unknown	++		
PSSP7_019	*gene 6* – exonuclease	+++		
PSSP7_019A	unknown	++		
PSSP7_020	*nrd* – ribonucleotide reductase	+++		
PSSP7_020A	unknown	+		
PSSP7_020B	unknown	+		
PSSP7_021	unknown	+		
PSSP7_022	unknown	++		
PSSP7_023	unknown	++	+++	
PSSP7_024	*gene 8* – head-to-tail connector	+++	++++	
PSSP7_025	*gene 9* – capsid assembly protein (scaffolding protein)	+++	+	
PSSP7_026	*hli* – high-light inducible protein	+		
PSSP7_027	*psbA* – D1 photosystem II reaction center protein	++		
PSSP7_028	Unknown			Cluster 3
PSSP7_029	*gene 10* – capsid protein	++++	++++	
PSSP7_030	*gene 11* – tail tubular protein A	+	++++	
PSSP7_031	*gene 12* – tail tubular protein B	++++	++++	
PSSP7_032	unknown (putative *gene 13?)*			
PSSP7_033	unknown (putative *gene 14*? – internal core protein)	++	++++	
PSSP7_034	*gene 15* – internal core protein	+++	++++	
PSSP7_035	*gene 16* – internal core protein	++++	++++	
PSSP7_036	*gene 17* – tail fiber	++++	++++	
PSSP7_037	unknown	++	+	
PSSP7_038	unknown	+	+++	
PSSP7_039	unknown	+	++++	
PSSP7_040	unknown	+		
PSSP7_041	unknown		+	
PSSP7_042	unknown		+	
PSSP7_043	unknown			
PSSP7_044	unknown	+		
PSSP7_045	unknown			
PSSP7_046	unknown		+++	
PSSP7_047	unknown			
PSSP7_048	unknown		++	
PSSP7_049	possible endonuclease			
PSSP7_050	unknown	++	++++	
PSSP7_051	*gene 19* – DNA maturase	+		
PSSP7_052	unknown			Cluster 2
PSSP7_053	unknown			
PSSP7_054	*talC* – transaldolase family protein	+++		

Supplementary Table 2: Promoter analyses: predictions and experimental detection of 5' ends upstream of the denoted ORF. Promoter analysis: nd – not determined. None = tested but no 5' end found. Processed 5' end = product found in both TAP- and TAP+ treatments. Motif = conserved motif found in proximity of processed 5' end.

ORF ID	Product	Bioinformatic Predictions	Experimental detection of 5' ends
PSSP7_001	unknown	Bacterial -10 box: 52..57	Processed 5' end=113 motif=91..116; Processed 5' end=44686; motif=44664..44689
PSSP7_002	unknown		nd
PSSP7_003	unknown	Terminator: 1675..1691	Processed 5' end=1672; motif=1650..1675
PSSP7_004	unknown		Same site as per PSSP7_003
PSSP7_005	unknown		nd
PSSP7_006	unknown		nd
PSSP7_007	unknown	Bacterial -10 box (x2): 2591..2596; 2629..2634	nd
PSSP7_008	unknown		Nothing conserved; Processed 5' ends (5x)=2444; 2446; 2450; 2451; 2459; Nothing conserved; Processed 5' ends (3x)= 2724; 2725; 2726
PSSP7_009	unknown		nd
PSSP7_010	unknown		nd
PSSP7_011	*gene 0.7* - MarR transcriptional regulator	Bacterial -10 box (x3): 3566..3571; 3577..3582, 3585..3590	none
PSSP7_012	*int* - phage related integrase		nd
PSSP7_013	*gene 1* - RNA polymerase	Terminator: 5005..5028 Bacterial -10 box: 5034..5039	Bacterial tis=5045 Non-Processed (x2) 5' ends=5032; 5060
PSSP7_014	*gene 2.5* - ssDNA binding protein		none
PSSP7_015	*gene 3* - endonuclease		nd
PSSP7_016	*gene 4* - primase/helicase		nd
PSSP7_017_018	*gene 5* - DNA polymerase		Nothing conserved; ; Processed 5' ends (3x)=9690; 9692; 9695
PSSP7_019	*gene 6* - exonuclease		none
PSSP7_019a	unknown		nd
PSSP7_020	*nrd* - ribonucleotide reductase domain	Possible -10 box (13160..13165), identical to that found experimentally for rpl21[40].	Nothing conserved; Processed 5' end=12818; Non-Processed 5' end=12928
PSSP7_020a	unknown		nd
PSSP7_021	unknown		nd
PSSP7_022	unknown		nd
PSSP7_023	unknown		nd
PSSP7_024	*gene 8* - head-to-tail connector		none
PSSP7_025	*gene 9* capsid assembly protein		nd
PSSP7_026	*hli* - high-light inducible protein		none
PSSP7_027	*psbA* - D1 photosystem II reaction center protein		none
PSSP7_028	unknown	Bacterial -10 box: 19430..19435	none (no signal found further upstream of Processed 5' end: 19749)
PSSP7_029	*gene 10* - capsid protein		Processed 5' end: 19749 motif: 19725..19750
PSSP7_030	*gene 11* - tail tubular protein A	Terminator: 21031..21046	none
PSSP7_031	*gene 12* - tail tubular protein B		nd
PSSP7_032	Unknown (*gene 13??*)	Terminator: 24626..24650	none
PSSP7_033	unknown (*gene*14??)		nd
PSSP7_034	*gene 15* - internal core protein		nd
PSSP7_035	*gene 16* - internal core protein		nd
PSSP7_036	*gene 17* - tail fiber		none
PSSP7_037	unknown		nd
PSSP7_038	unknown		nd
PSSP7_039	unknown		nd
PSSP7_040	unknown		nd
PSSP7_041	unknown		nd
PSSP7_042	unknown		nd
PSSP7_043	unknown		nd
PSSP7_044	unknown		nd

PSSP7_045	unknown		nd
PSSP7_046	unknown		nd
PSSP7_047	unknown		nd
PSSP7_048	unknown		nd
PSSP7_049	possible endonuclease		nd
PSSP7_050	unknown	Terminator: 39402..39417	none
PSSP7_051	*gene 19* - DNA maturase	Terminator: 41165..41176	none
PSSP7_052	unknown	Bacterial -10 box (x2): 42978..42983; 42988..42993	none
PSSP7_053	unknown		nd
PSSP7_054	*talC* - transaldolase		nd
		Terminator: 44063..44075	

Supplementary Table 3: Upregulated *Prochlorococcus* MED4 genes determined from microarray analysis. Fold change (infected/control) with time (h) after infection. Positive and negative values indicate an increase and decline in transcript levels respectively. Significant increases in fold change are shown in blue and the level of significance is shown: * for q<0.05; ** q<0.01; ***q<0.001.

ORF – gene name, possible product and function				Fold	Change	(inf/ctrl)			
TRANSCRIPTION GROUP 1	**0 h**	**1 h**	**2 h**	**3 h**	**4 h**	**5 h**	**6 h**	**7 h**	**8 h**
PMM0549 – *csoS1* carboxysome shell protein 1, carbon fixation	1.70	1.31**	-1.22	-1.59	-1.76	-2.07	-2.03	-2.16	-2.58
PMM0550 – *rbcL* rubisco large subunit, carbon fixation	1.79	2.01***	1.38**	1.18*	-1.41	-1.69	-1.62	-2.00	-2.05
PMM0551 – *rbcS* rubisco small subunit, carbon fixation	1.63	1.85***	1.36**	1.17*	-1.43	-1.67	-1.62	-2.04	-2.01
#PMM0815/PMM1396 – *hli19/09* high-light inducible stress response protein	1.15	1.23*	-1.27	-1.45	-1.25	-1.38	-1.45	-1.58	-1.39
#PMM0816/PMM1397 – *hli18/08* high-light inducible stress response protein	1.09	1.29**	-1.12	-1.26	-1.26	-1.50	-1.56	-1.64	-1.56
#PMM0817/PMM1398 – *hli17/07* high-light inducible stress response protein	1.16	1.35***	-1.17	-1.19	-1.24	-1.47	-1.46	-1.64	-1.60
#PMM0818/PMM1399 – *hli16/06* high-light inducible stress response protein	1.14	1.33**	-1.18	-1.24	-1.28	-1.53	-1.49	-1.66	-1.70
PMM0970 – *urtA* urea ABC transporter periplasmic binding protein	1.34	1.43***	1.01	-1.24	-1.51	-1.77	-1.59	-1.75	-1.52
PMM1135 – *hli14* high-light inducible stress response protein	1.24	1.26**	-1.17	-1.26	-1.54	-1.76	-1.73	-1.88	-1.94
PMM1483 – *rpoC2* RNA polymerase subunit, transcription	1.01	1.26*	-1.02	-1.28	-1.49	-1.59	-1.61	-1.67	-1.67
PMM1536 – *rps11* ribosome small subunit protein 11, translation	1.35	1.19*	-1.09	-1.35	-1.80	-1.94	-1.95	-2.05	-2.09
PMM1544 – *rpl6* ribosome large subunit protein 6, translation	1.02	1.23*	-1.05	-1.44	-1.70	-1.74	-2.04	-1.86	-1.97
PMM1545 – *rps8* ribosome small subunit protein 8, translation	1.07	1.21*	-1.19	-1.39	-1.68	-1.86	-1.93	-1.85	-1.90
PMM1546 – *rpl5* ribosome large subunit protein 5, translation	-1.06	1.33**	-1.09	-1.33	-1.82	-1.94	-2.22	-1.94	-1.94
PMM1549 – *rps17* ribosome small subunit protein 17, translation	-1.13	1.22*	-1.17	-1.40	-1.97	-2.09	-2.15	-2.10	-1.98
PMM1629 – *rpoD type II* alternative sigma factor, transcription	1.10	1.61***	-1.09	-1.34	-1.38	-1.74	-1.80	-1.84	-1.92
TRANSCRIPTION GROUP 2	**0 h**	**1 h**	**2 h**	**3 h**	**4 h**	**5 h**	**6 h**	**7 h**	**8 h**
PMM0014 – *dus* tRNA dihydrouridine synthase, RNA modification	1.20	-1.28	1.23*	1.21*	1.44*	1.48**	1.06	1.49**	1.21
PMM0030 – unknown	-1.10	-1.15	1.09	1.66**	1.41*	1.08	-1.41	1.00	1.00
PMM0334 – unknown	1.11	-1.49	-1.07	1.05	1.20	1.26*	-1.11	1.04	-1.17
#PMM0368 – unknown	1.02	-1.05	1.21*	1.70***	1.85**	1.61***	1.68***	1.70***	1.32
PMM0426 – *sun* tRNA and rRNA methyltransferase, RNA modification	1.09	-1.66	-1.48	-1.20	1.48*	1.41**	1.23	1.81***	1.49*
#PMM0684 – unknown (homologous to PMM0819 and PMM1134)	1.01	-1.06	1.27*	1.58***	1.35	1.16	1.08	1.07	-1.16
#PMM0685 – unknown (homologous to PMM1427)	-1.05	-1.36	1.77***	2.65***	2.37**	2.01***	1.44*	1.49*	1.39*
#PMM0686 – *clpS-like* protease adaptor, protease inhibition and redirection	-1.07	-1.48	3.91***	8.95***	14.00***	11.71***	8.55***	10.82***	8.75***
#PMM0819 – unknown (homologous to PMM0684 and PMM1134)	1.28	-1.01	1.52***	2.16***	2.32**	1.84***	1.66***	1.57***	1.21
PMM0830 – *DHPS-like* folate biosynthesis, nucleotide & amino acid synthesis	1.16	-1.84	-1.69	-1.38	1.38*	1.36**	1.10	1.18	-1.12

PMM0936 – *umuD* SOS response to DNA damage	1.01	-1.55	**1.31***	**1.39*****	**1.75****	**1.53*****	1.15	**1.43****	1.10
PMM1114 – unknown	1.26	-1.81	-1.05	-1.18	1.22	**1.29***	1.05	1.04	1.02
PMM1115 – *crtH*, phytoene dehydrogenase, secondary metabolite biosynthesis	1.01	-1.42	-1.15	-1.07	1.26	**1.17***	-1.08	1.00	-1.14
PMM1187 – AAA ATPase family, protein turnover, stress response	1.01	-1.39	1.09	**1.21***	**1.56***	**1.45****	1.00	**1.27***	1.15
#PMM1201 – dTDP-D-glucose 4,6-dehydratase, cell envelope biogenesis	1.27	-1.63	-1.43	-1.16	1.22	**1.27****	1.02	1.17	-1.08
#PMM1248 – unknown	1.08	-2.27	-1.76	-1.43	**1.37***	**1.60*****	1.12	**1.33***	1.25
PMM1284 – *phoH-like* phosphate stress ATPase	1.13	-1.34	1.06	1.11	**1.80****	**1.56*****	1.12	**1.26***	1.07
#PMM1403 – HNH nuclease domain, site-specific endonuclease	1.11	-1.88	-1.56	-1.60	1.24	**1.45****	1.11	**1.56****	**1.51***
#PMM1426 – unknown	1.47	-2.50	-2.05	-1.82	1.16	**1.49****	1.18	**1.31***	1.36
#PMM1427 – unknown (homologous to PMM0685)	1.00	-1.20	**1.23***	**1.71*****	**1.93****	**1.74****	**1.52****	**1.54****	1.35
PMM1428 – unknown	1.02	-1.27	-1.03	1.03	**1.60****	**1.50****	1.16	**1.32***	1.04
PMM1501 – *me* RNase E, mRNA degradation	1.01	-1.16	**2.44*****	**3.51*****	**3.43*****	**2.91*****	**1.83*****	**2.09*****	**1.72***
PMM1502 – *mhB* RNase HII, DNA replication and repair	1.13	-2.16	**1.47***	**2.25*****	**3.55*****	**2.91*****	**1.64****	**2.52*****	**2.09****
PMM1517 – unknown	-1.06	-1.58	-1.32	1.03	**1.39***	**1.31***	1.10	1.06	1.15
PMM1529 – *prfA* peptide release factor, translation	1.31	-1.11	1.10	**1.52*****	**1.67****	**1.29****	-1.04	1.17	-1.10

#Genes found in genome islands as per Coleman et al.[6].

126

Supplementary Table 4: High-light inducible genes (*hli*) in *Prochlorococcus* MED4 and their expression patterns during exposure to environmental stressors.

MED4 Gene IDs	High Light Stress	Nitrogen Stress	Phage Infection
PMM0093 - *hli01*			
PMM0064 - *hli02*			
PMM1482 - *hli03*			
*PMM1118 - *hli04*	+		
*#PMM1404 - *hli05*	+		
*#PMM0818/PMM1399 - *hli06/hli16*	+		+
*#PMM0817/PMM1398 - *hli07/hli17*	+		+
*#PMM0816/PMM1397 - *hli08/hli18*	+		+
*#PMM0815/PMM1396 - *hli09/hli19*	+		+
*#PMM1390 - *hli10*		+	
*#PMM1385 - *hli11*	+		
*#PMM1384 - *hli12*	+		
PMM1317 - *hli13*			
*PMM1135 - *hli14*	+		+
*#PMM1128 - *hli15*	+	+	
PMM0471 - *hli20*			
*#PMM0690 - *hli21*	+	+	
*#PMM0689 - *hli22*	+	+	

*Clusters with phage *hli* genes as per Lindell & Sullivan et al.[5]; #Found in genome islands as per Coleman et al.[6]. High-light stress[46], Nitrogen stress[23], Phage infection (this study).

Supplementary Table 5: Comparison of array normalization methods to RT-PCR. Significance is assigned to differentially expressed genes determined by RT-PCR (p-values) normalized to *rnpB* and microarray analysis (q-values) normalized using different methods (Quantile RMA; Hyb Ctrl;.Golden Spike (GS) regular[32]; GS without 2nd Loess)

Expression Pattern	ORF *gene*	Time	RT-PCR	RMA	Hyb Ctrl	GS w/ 2nd loess (regular)	GS w/o 2nd Loess
Unchanged	PMM_rnpB *rnpB*	0	0.528	0.912	0.759	0.991	0.978
		1	0.530	0.818	0.674	0.419	0.795
		3	0.938	0.805	0.158	0.614	0.550
		4	0.878	0.489	0.712	0.884	0.986
		8	0.970	0.756	0.281	0.627	0.903
DownRegulated	PMM0496 *rpoD*	0	0.171	0.529	0.231	0.995	0.998
		4	0.000	0.003	0.470	0.539	0.123
		8	0.001	0.003	0.043	0.966	0.100
	PMM0627 *pcb*	0	0.422	0.337	0.150	0.655	0.769
		4	0.004	0.001	0.449	0.572	0.208
		8	0.000	0.000	0.007	0.152	0.048
	PMM1309 *ftsZ*	0	0.034	0.405	0.072	0.813	0.803
		4	0.000	0.008	0.602	0.703	0.113
		8	0.000	0.000	0.003	0.453	0.011
	PMM1629 *rpoD type II* (up- and down-regulated)	0	0.195	0.231	0.123	0.889	0.895
		1	0.388	0.000 up	0.636	0.000 up	0.044
		3	0.000 dn	0.001 dn	0.005	0.457	0.057
		8	0.000 dn	0.001 dn	0.054	0.991	0.042
UpRegulated	PMM0550 *rbcL* (up- and down-regulated)	0	0.041 up	0.974	0.000 up	0.022 up	0.016 up
		1	0.066	0.000 up	0.283	0.000 up	0.000 up
		3	0.317	0.010 up	0.021	0.000 up	0.170
		8	0.269	0.000 dn	0.051 dn	0.809	0.050
	PMM0684 unknown	0	0.095	0.999	0.200	0.992	0.989
		3	0.049	0.000	0.000	0.000	0.005
		4	0.001	0.088	0.223	0.000	0.171
		8	0.075	0.269	0.154	0.000	0.702
	PMM0686 *clpS-like*	0	0.467	0.337	0.231	0.801	0.624
		4	0.006	0.000	0.000	0.000	0.000
		8	0.000	0.000	0.000	0.000	0.000
	PMM0819 unknown	0	0.599	0.719	0.101	0.373	0.476
		4	0.000	0.002	0.035	0.000	0.003
		8	0.013	0.164	0.014	0.000	0.230
	PMM0936 *umuD*	0	0.439	0.193	0.416	0.889	0.921
		1	0.276	0.000 dn	0.041	0.251	0.038 dn
		3	0.001	0.000	0.005	0.006	0.059
		4	0.005	0.005	0.101	0.003	0.075
		8	0.001	0.497	0.350	0.459	0.843
	PMM1284 *phoH-like*	0	0.002	0.486	0.105	0.707	0.765
		4	0.002	0.002	0.037	0.003	0.111
		8	0.738	0.569	0.336	0.256	0.754
	PMM1501 *rne*	0	0.451	0.297	0.403	0.993	0.988
		3	0.025	0.000	0.000	0.000	0.000
		8	0.026	0.019	0.036	0.027	0.160

p determined from 2-tailed t-test for RT-PCR results and q determined using Cyber-T and Q-value for microarray results.

Supplementary Table 6: Mass spectrometric detection of previously unannotated proteins from the P-SSP7 phage genome. Detected tryptic peptides are in bold and underlined. The entire ORF inferred from these peptide detections are shown. For PSSP7_033, detection of the peptide suggests an N-terminal extension of the previously annotated protein (shown in italics).

GENE ID	PROTEIN SEQUENCE
PSSP7_19A	MTTRKK**NQSFGPPPPITK**LTTEQDFKLRQLEILLSKPETRKEDIAIVMIALQE QAFVLSNCIKNLIEKWPKPPTTTDPR**TTNEVPLMFGILLETK**DSDFTSET
PSSP7_20A	MKYLGEVVRTVTVPAFYTLILITPILLTSCKSKDKNSHGLNDVWTSPENSGLI QKLEQR**KQLYKELLGETSGSTK**
PSSP7_20B	**METESIQTSVLR**FTCPHAERASYSTLICQPVVSATYEKVCVKVCQICASSIV GQGLKNLESILHQISTGKLDSDS
PSSP7_033	MCLGAAAKAANENARRRYKYENERRER***NWMQTMSIYNAQK****VKYDEDVQ NAGLAQAQVKTDQQEAMDLARGEAQIKYAELFRKLLNDSTYGKLVASGQT GQSTRRRATMDYAKYGRDVSDIARRLTLNDRELARKSSEQISKYKQFKDE AFAKVAFQPIPDVAPPQPVMRNVGAEAFMGALSIASNVATMGGQSGFGW WGG*

Supplementary Table 7: Primers used for RT-PCR verification of microarray results and normalization methods for representative phage and host genes.

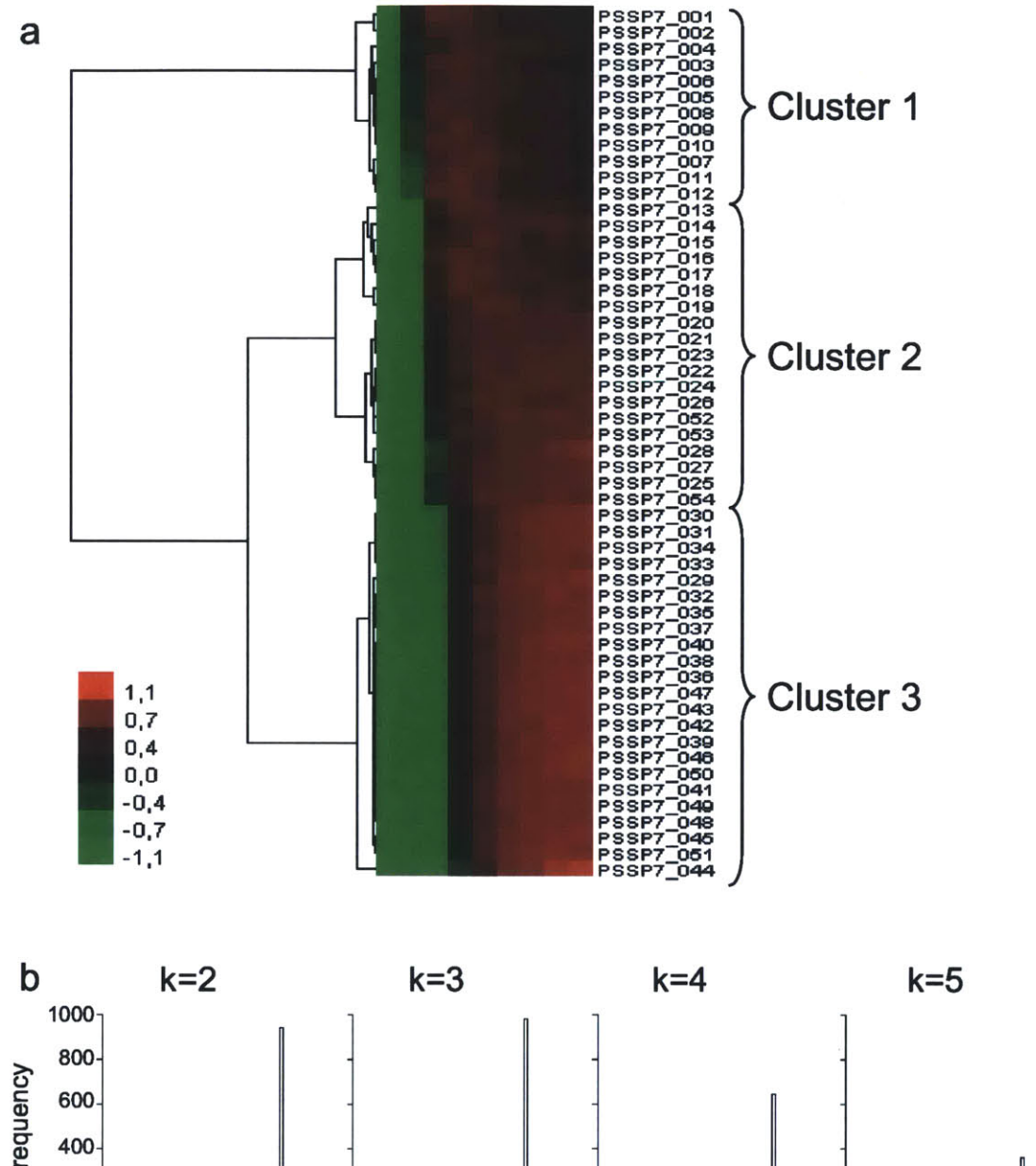
Target ORF *gene* – product	Primer Direction	Primer Sequence 5' – 3'
Phage P-SSP7 genes		
PSSP7_001 Unknown	F R	CCAAGCCAAAGGCTACACAT GCATCCCTTGATTCATTGCT
PSSP7_003 Unknown	F R	ATGGTTCACTTCCTAACCAAGC CCCCCTTACCCATAGGTGTT
PSSP7_013 *gene 1* – RNA polymerase	F R	CGACTATGGAGGAGCGGTTA GTCTGCTGCTTCCCAATCTC
PSSP7_017 *gene 5* – DNA polymerase	F R	AAACACTTCCGCCCTTACCT CTGCAACGAAAGGGAATTGT
PSSP7_020 *nrd* – ribonucleotide reductase	F R	TTGTGCAAGCTCCATAGTCG GCCTTACCAAACTCGGCATA
PSSP7_027 *psbA* – D1 photosystem II protein	F R	CTCTGCTATGCACGGAAGTT GCAGATTCCCATGGAGGTAA
PSSP7_029 *gene 10* – capsid protein	F R	GGCTTCCAGCATGAAACAAT TGGTCTTCTCTGCAACTGGA
PSSP7_054 *talC* – transaldolase family protein	F R	TGGTCGAAAATACGGAGAGG TACGTAGCACCAGCATGAGC
Host *Prochlorococcus* MED4 genes		
PMM_rnpB *rnpB* – RNA of RNase P	F R	TTGAGGAAAGTCCGGGCTC GCGGTATGTTTCTGTGGCACT
PMM0496 *rpoD* – principle RNA polymerase sigma factor	F R	AATCAGAGCTGCCGAAAATA TGATCTGCTATCGCTCGTGT
PMM0550 *rbcL* – rubisco large subunit	F R	CCTGAATATGTCCCCCTCGA CCGCTGCTGCAACTTCTTCT
PMM0627 *pcb* – chlorophyll a/b binding protein	F R	TCATGTCGCTCATGCAGGG GACCCATTGGGACACTGGG
PMM0684 Unknown	F R	CGCAAGGCAGCTTTTTAATC TCCATGTTTCAAACGCAGAG
PMM0686 *clpS-like* – protease adaptor	F R	CAGTTGTAGATCCAAAGACAACG CAAGACAATTTGCTACGTGTTCA
PMM0819 Unknown	F R	CCCAAGTGGTTGGCTTCTTA ATCCCAGGCTTTTTCCAAAT
PMM0936 *umuD* – SOS response to DNA damage	F R	GTGATTCGGTCTCAGCAGGT TTCTCCATCTATCATCGCAATA
PMM1284 *phoH-like* – phosphate stress induced ATPase	F R	GTTTGTGCCGCCAGATTATT TGCTAATGGTGCGACTTCAA
PMM1309 *ftsZ* – cell division protein	F R	AATGACTGAAGCTGGCACTGC ACTATTCATTGCGGCTTGAGC
PMM1501 *rne* – RNase E	F R	AACCGCCTAGCACAGGATTA TGCTTTTTCGAGAGCGATTT
PMM1629 *rpoD type II* – alternative sigma factor	F R	GAGTTGCCCGAAGATGATGT ACATTGGCTCATCTCCATCC

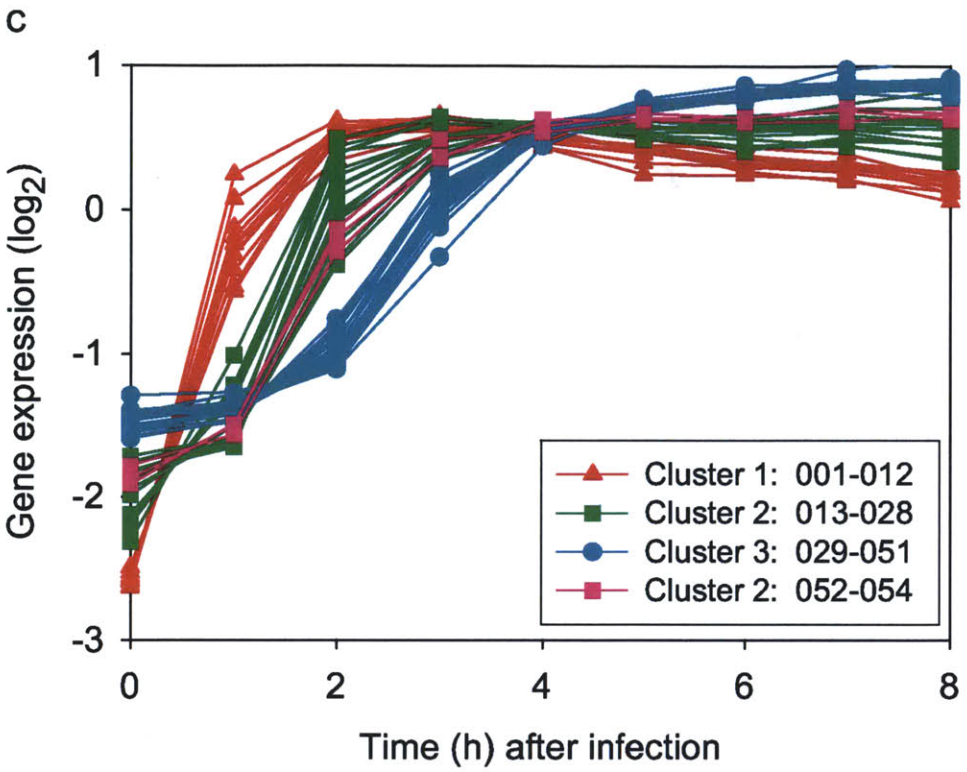
Supplementary Table. 8. Primers used in 5' RACE analysis for phage P-SSP7 and host *Prochlorococcus* MED4 genes. Primers used for reverse transcription are designated by "rt", whereas those used for nested or second nested PCR are designated by "nest" and "nest2" respectively in the oligonucleotide name. "up" – the primer was designed upstream of the gene.

Gene	Oligo. Name	Oligonucleotide Sequence (5'–3')	Length
Phage P-SSP7 genes			
PSSP7_001	P001rtREV	TCTCCATAATTGACCCGCTT	20
	P001nestREV	CTTAATAAAGTCAGTCAGTCCATCCCAGTCAGT	33
	P001nest2rev	TGATGGGAGAGGAATTGAACCTCTCGAT	28
PSSP7_003	P003nestREV	CCTTCTTACCGAATTTTCCTATGGTGTACTTAG	33
	P003rtREV	TCTCCCCCTTACCCATAG	18
PSSP7_004	P004nestREV	GGAATCGTTTACAGGGAATAACTCAGGCTCG	31
	P004rtREV	TTGTGCTTGGGTTTCTCTCA	20
PSSP7_008	P008nestREV	CGAAGGCTTCATAAGGCATCGAAGGAAAG	29
	P008rtREV	GGTATCTGATAGCCATAAATCTT	23
PSSP7_011	P011nestREV	CAGTCAGTATTACGGCTACCACTTGCGC	28
	P011rtREV	TCAATCTATGAAACTCACTGAG	22
PSSP7_013	P013rtREV	TCTTTACGTGGGGAGAATAG	20
	P013nestREV	GGTTATCTTTGCAGTTATAGCTCCTTGCGATTCTG	35
PSSP7_014	P014nestREV	CTATAAACTTACCTTCTTCTACCTCCTCCCATG	33
	P014rtREV	CTGGAGGACGCTTATCTTC	19
PSSP7_017	P017nestREV	CAGCATGGGTGAGCCAATGCAGAGC	25
	P017rtREV	ACAGGTAAATCGTAGCCAATAAT	23
PSSP7_019	P019nestREV	CTTAAGTTCCTGTATGACACGTTTGTATCCAC	32
	P019rtREV	CATCTGCTTCTAGAGTATCTC	21
PSSP7_020	P020rtREV	GTCCTGATGCAACGAGAG	18
	P020nestREV	CCCTTATTTGTTCTTGTTCCCGCCGGTC	28
PSSP7_020 up	P020nest2REV	CCGAGGTGAAATCCGAGTCCTTGGTC	26
	P020rtREV	ACCCTGCTCTGCATGTGT	18
PSSP7_024	P024nestREV	GGAGGTAGAACTGCGAGCATGAGCTTC	27
	P024rtREV	CCTAACTTGTCATCACGTAC	20
PSSP7_026	P027nest2REV	CAGCCATTAAATCTTTCTGCTTCTGGTGACATTAG	35
PSSP7_027	P027nestREV	CCTATTGCATTGGAGCTTGGAACTACTGC	29
	P027rtREV	CGGCTTCCCAGATCGG	16
PSSP7_029 up	P029nest2REV	GGACTCGCACCTCCCTGATCGG	22
PSSP7_029	P029rtREV	GCCTTGTCAGCATTACCC	18
	P029nestREV	GGAAGGAACTTGTCATACGACCTGTGTAG	29
PSSP7_030	P030nestREV	GCCATCCTTCACCCTGTACATCTTTGTTTG	30
	P030rtREV	TCTGGGGTAACAAGTACATG	20
PSSP7_032	P032nestREV	CATCTTCCTGTACGCCGGCTACTCC	25
	P032rtREV	CGGGTGTACATAGCATCC	18
	P032nest2rev	CTCCATAAGCGGCACATAGTGGGATTACC	29
PSSP7_036	P036nestREV	GTCGAGTTCAACTTTGACATCGACATTCGC	30
	P036rtREV	GGTGTAGTCATTATTTGTTTGAC	23
PSSP7_050	P050nestREV	GTCCATTATTTCAGGGTTAGCTTTTTGTGCAGG	33
	P050rtREV	ACCTTTCCATCTCATTTTAGAGA	23
PSSP7_051	P051nestREV	GGCTTGAATCTGGAGTCTCTTGGGTCC	27
	P051rtREV	CCAAGATTTACCAACACCTC	20
PSSP7_052	P052rtREV	CCTTTGCGTTAAAGATGCTG	20
	P052nestREV	CTTCCATCTCATCCAGGTAGTCCTCAATAGC	31
Host MED4 genes			
PMM0368	PMM0368nestREV	CTATTGCCCAACCATTAGCTCCCTGAGG	28
	PMM0368rtREV	GCTGACACCACTGCCAAA	18
PMM0684	PMM0684nestREV	CTGCCTTGCGGGTTAAGTATCCAGCC	26
	PMM0684rtREV	TTGTAAATAAGAGATGGGTATAAAC	25
PMM0819	PMM0819nestREV	TGACTTGATGACTTGATTGCTGAGGGCTC	29
	PMM0819rtREV	CATTTATCCCAGGCTTTTTCC	21
PMM1500	PMM1500nestREV	CCATCCATATCCTTGCTCTGGGCCG	25
	PMM1500rtREV	GATGAGATTTGACACCCTC	19

PMM1501	PMM1501nestREV	CTATAAAGGCAGCATCAATACCTGGTAGGAC	31
	PMM1501rtREV	GGACCTAGATCTGATACATG	20

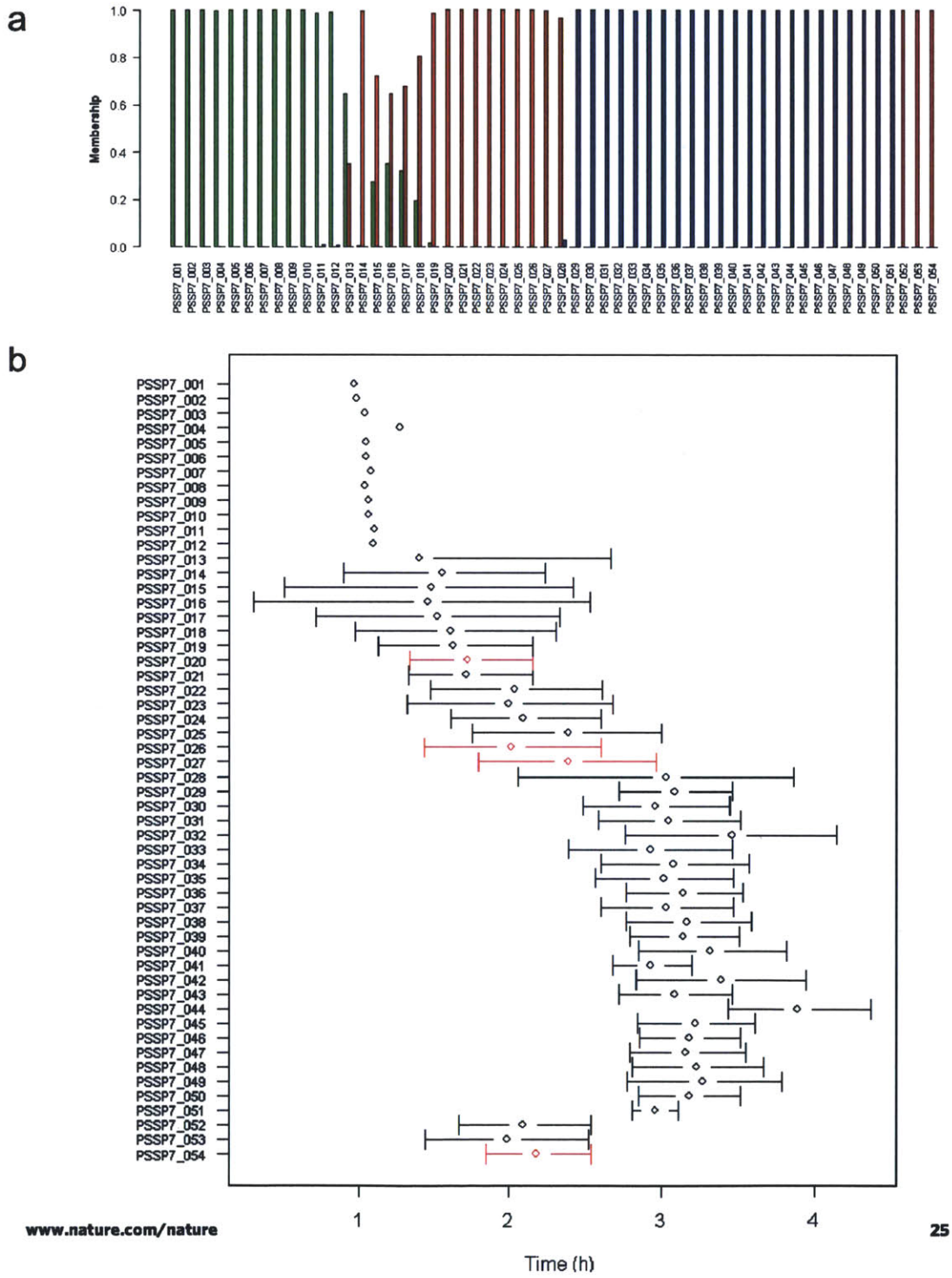
Supplementary Figure 1



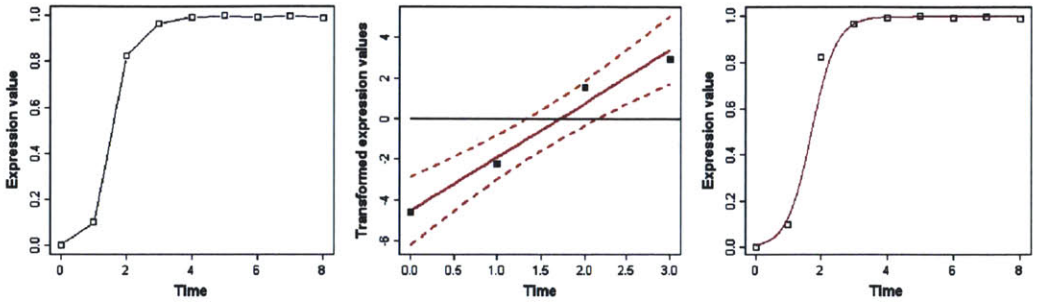


Supplementary Figure 1. Cluster analysis of phage gene expression profiles. (a) Hierarchical clustering of phage gene expression profiles, after standardization of logged data (mean expression equal to zero and standard deviation equal to one) was performed with average linkage and Pearson correlation. (b) Distribution of Jaccard coefficients derived from 1000 random independent resamplings of phage genes. The proportion of genes used for resampling was 0.7. The number of clusters (k) tested ranged from 2 to 5. Average linkage and Pearson correlation was used for the hierarchical re-clustering. For k=3 clusters, 982 out of 1000 Jaccard coefficients equaled 1 indicating that phage genes form three stable clusters. (c) Temporal profiles of the 3 clusters detected by hierarchical clustering. See Figure 2a for a representation of these temporal profiles after minimum-maximum normalization. Note that the last 3 genes in the genome cluster together with genes from cluster 2 and are transcribed prior to genes in cluster 3. See Suppl. Fig. 2 for statistical analysis of the significance of the clustering of all 4 'bacterial-like' genes in cluster 2. See Suppl. Table 1 for gene name and function for each ORF.

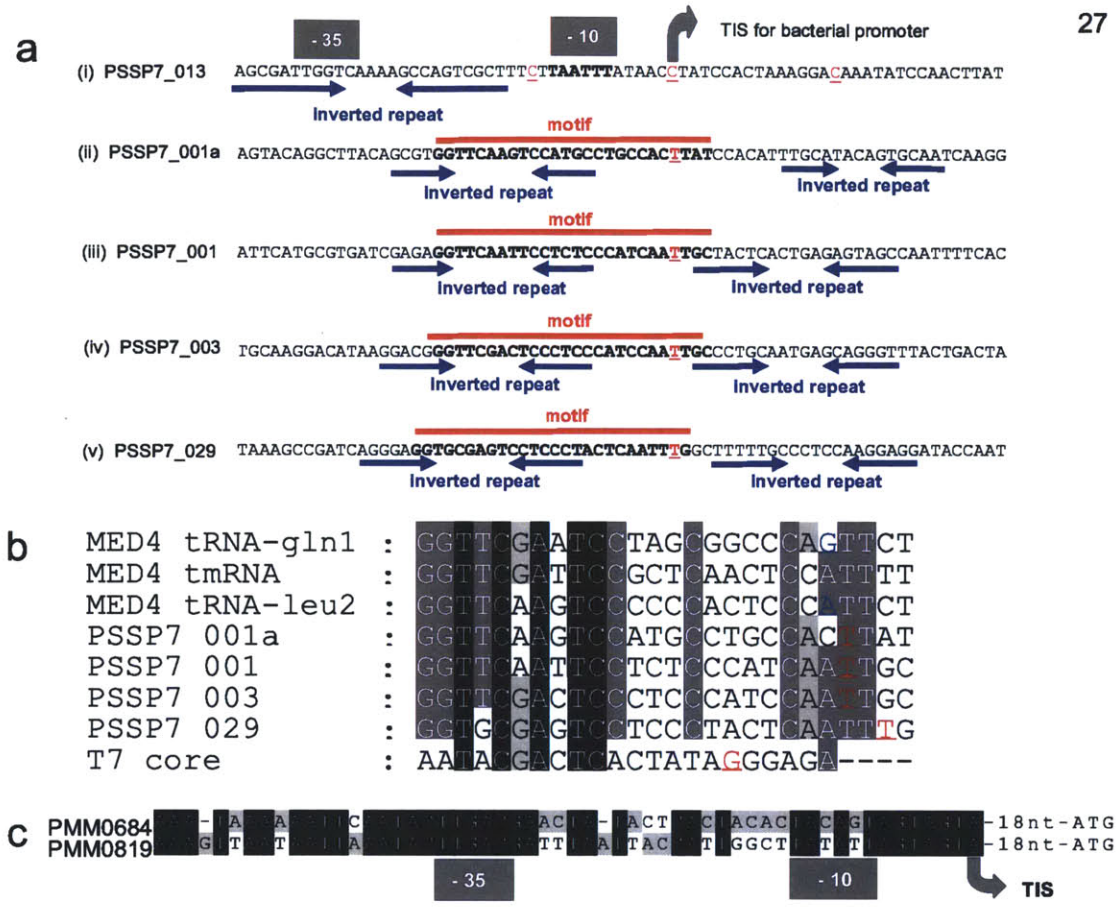
Supplementary Figure 2



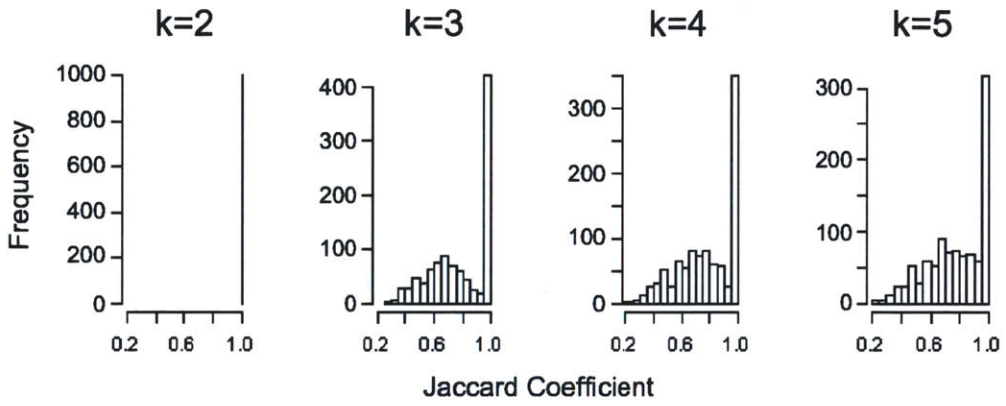
C



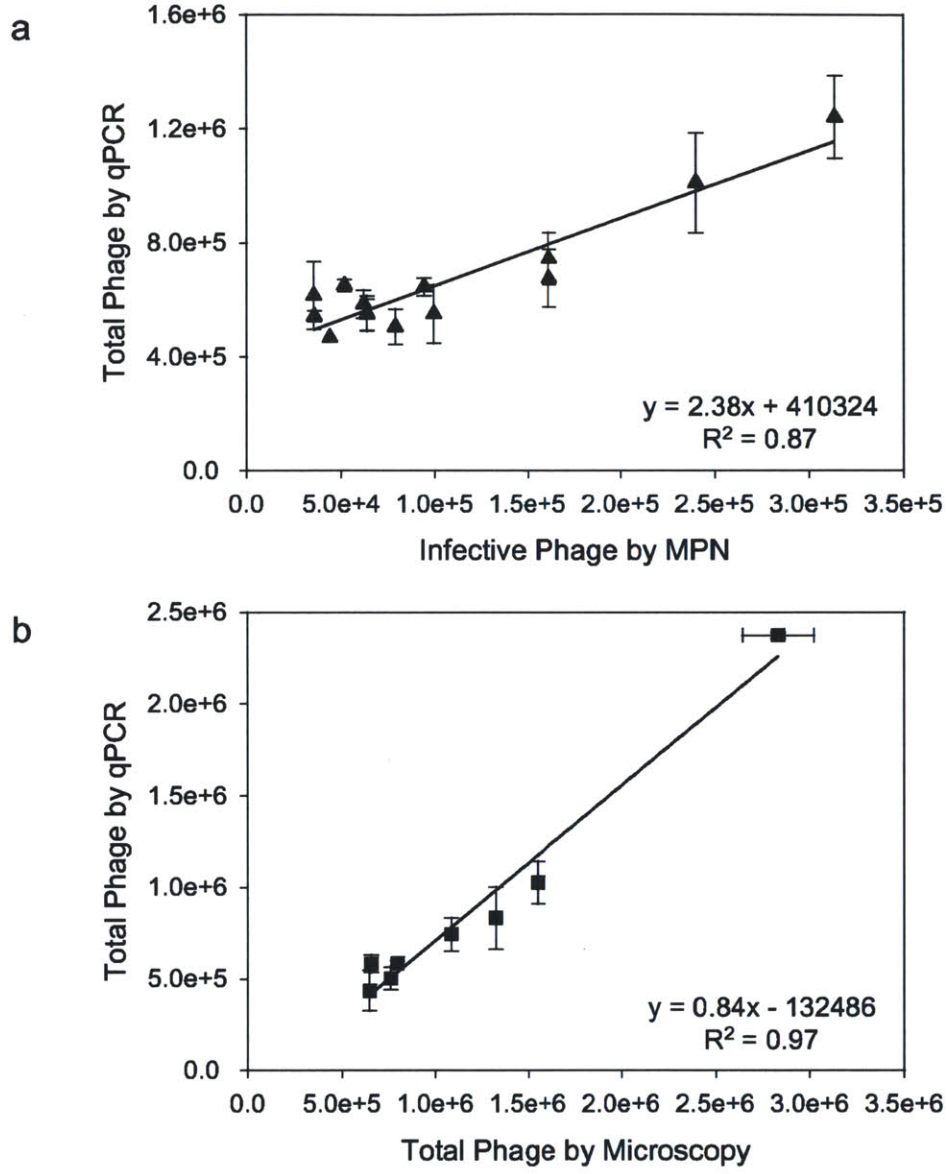
Supplementary Figure 2. Significance of the temporal coexpression of the last 3 genes of the genome with genes in cluster 2 and therefore of the 4 'bacterial-like' genes *nrd* (020), *hli* (026), *psbA* (027) and *talC* (054). (a) Cluster membership for genes in cluster 1 (green), cluster 2 (red) and cluster 3 (blue) were based on 10000 independent bootstrap samplings with replacement of expression values for the same gene. (b) 90% confidence intervals are shown for the switch time (t*) at which transcription of the phage genes went from being non-expressed to expressed – defined here as the time point at which 50% of maximal expression was reached. Confidence intervals for the 4 'bacterial-like' genes are shown in red. Note that no intervals could be derived for ORFs 1-12 due to immediate initiation of expression. (c) An example of the fitting procedure carried out (for *nrd*) to determine the confidence intervals shown in (b). The left panel shows the expression data (y) after normalization so that minimum expression equals zero and maximal expression equals 1. The middle panel shows the linear regression (solid red line) and the confidence intervals (dashed red lines) determined after the transformation y'=log(y/(1-y)). Note time t* is derived as the time point for which the regression line crosses y'=0 (set at half maximum expression). The regressed sigmoidal curve is shown in the right panel. The reliable assignment of the last 3 genes on the genome with genes in expression cluster 2 as well as the overlap of their confidence intervals provides strong evidence for the temporal coexpression of the 4 'bacterial-like' genes in cluster 2 despite their spatial separation on the genome.



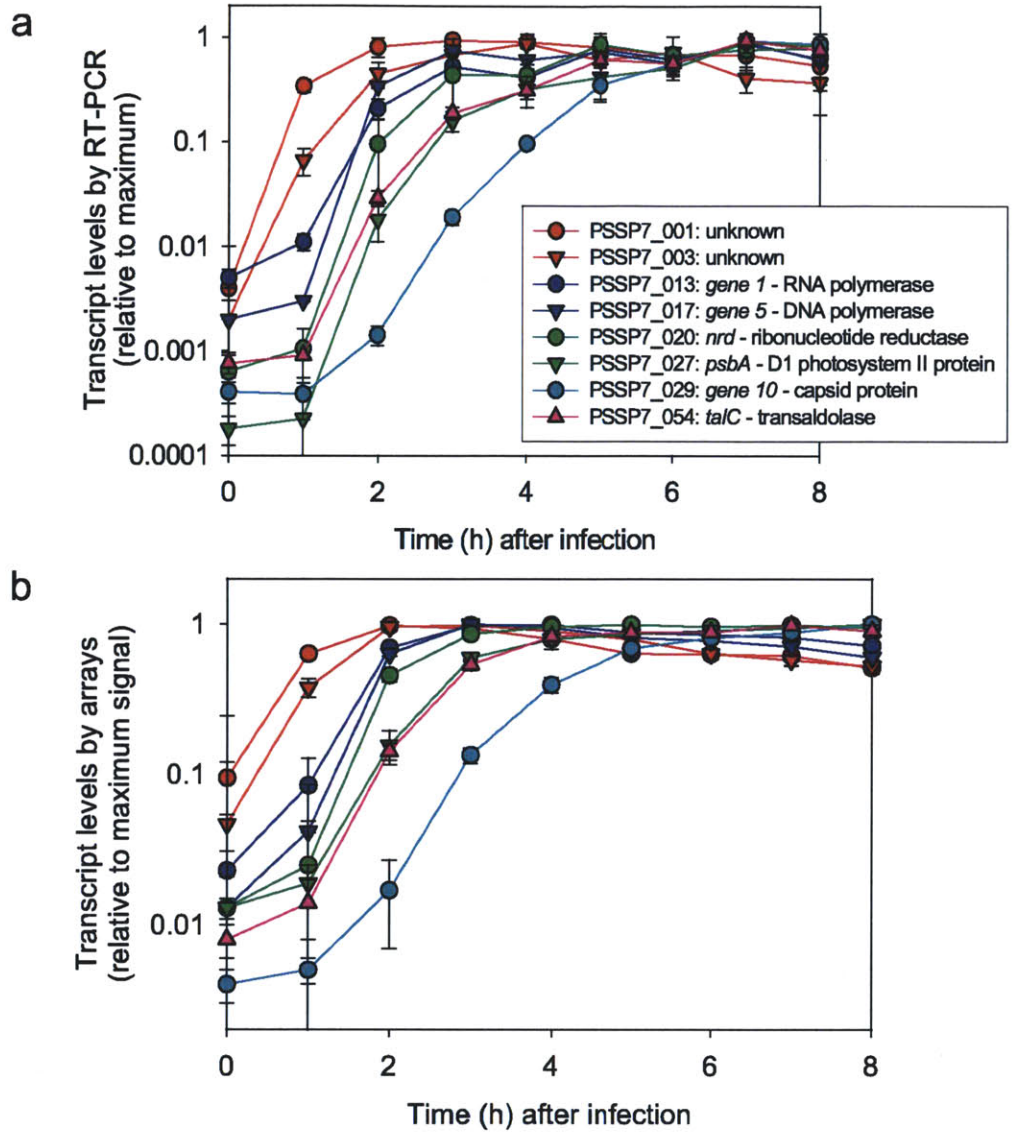
Supplementary Figure 3. RACE mapping of 5' transcript ends for phage and host genes and associated possible regulatory elements. (a) Experimentally mapped 5' ends for phage genes. A typical bacterial promoter was found upstream of the 5'ends of the cluster 2 gene PSSP7_013 coding the RNA polymerase (i). The 5' ends upstream of cluster 1 and 3 genes – ORF PSSP7_001 (2 transcripts), PSSP7_003 and PSSP7_029 – were found in both TAP+ and TAP- treatments (data not shown) suggesting that they are mature transcripts that have undergone post-transcriptional processing (ii to v). The nt at the 5' ends are shown in red and underlined. A 26 nt conserved motif (bold type face and marked with a red line above the sequence) was found in the region of these processed 5' ends, as were inverted repeats (blue arrows below the sequence). (b) Comparison of the conserved motif found upstream of processed transcripts in (a) shows that they have weak similarity to the T7 core promoter. A bioinformatic search for sequence similarity between this motif and the MED4 host genome revealed similarity to putative 3' cleavage sites of 2 tRNA genes and tmRNA, suggesting that the inverted repeats associated with this motif may serve as an RNase III recognition site in a similar fashion to that known for T7 with linked promoter and processing sites[47]. The red underlined nucleotide indicates the 5' end determined by RACE and for the T7 core promoter it indicates the known transcript start. The annotated 3' end of the tRNAs is blue and underlined. (c) Transcription initiation sites for two up-regulated MED4 homologous genes of unknown function (PMM0684 and PMM0819). The consensus bacterial regulatory elements (the -10 and -35 box) are shown upstream of the transcription initiation site.



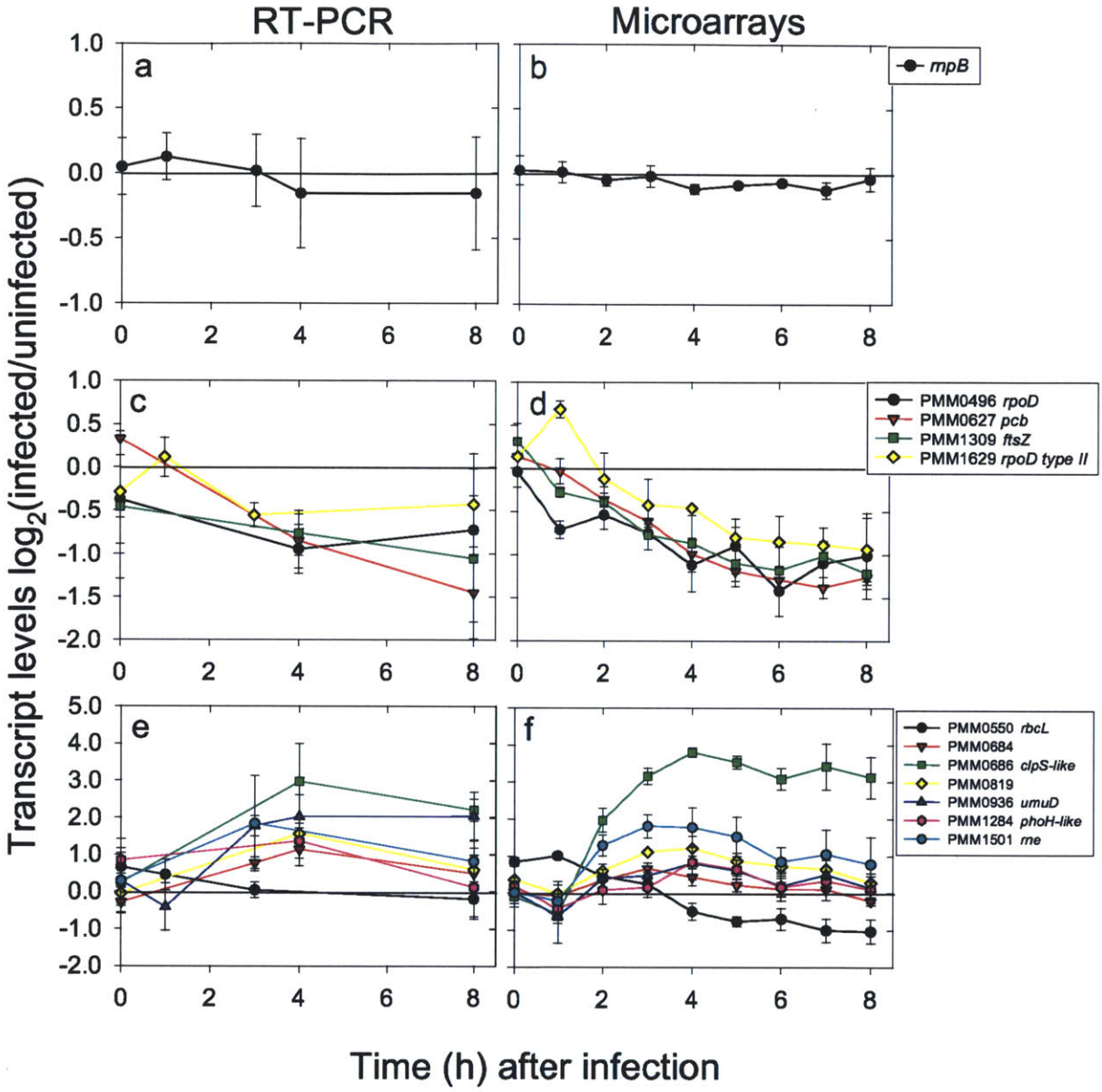
Supplementary Figure 4. Analysis of the number of stable clusters of upregulated host (MED4) genes. The distribution of Jaccard coefficents was derived from 1000 independent random samplings of upregulated host genes. The number of clusters (k) tested ranged from 2 to 5. The proportion of genes used for resampling was 0.7. For hierarchical re-clustering average linkage and Pearson correlation was used. For k=2 clusters the coefficients are concentrated at 1 indicating that upregulated host genes form two stable clusters.



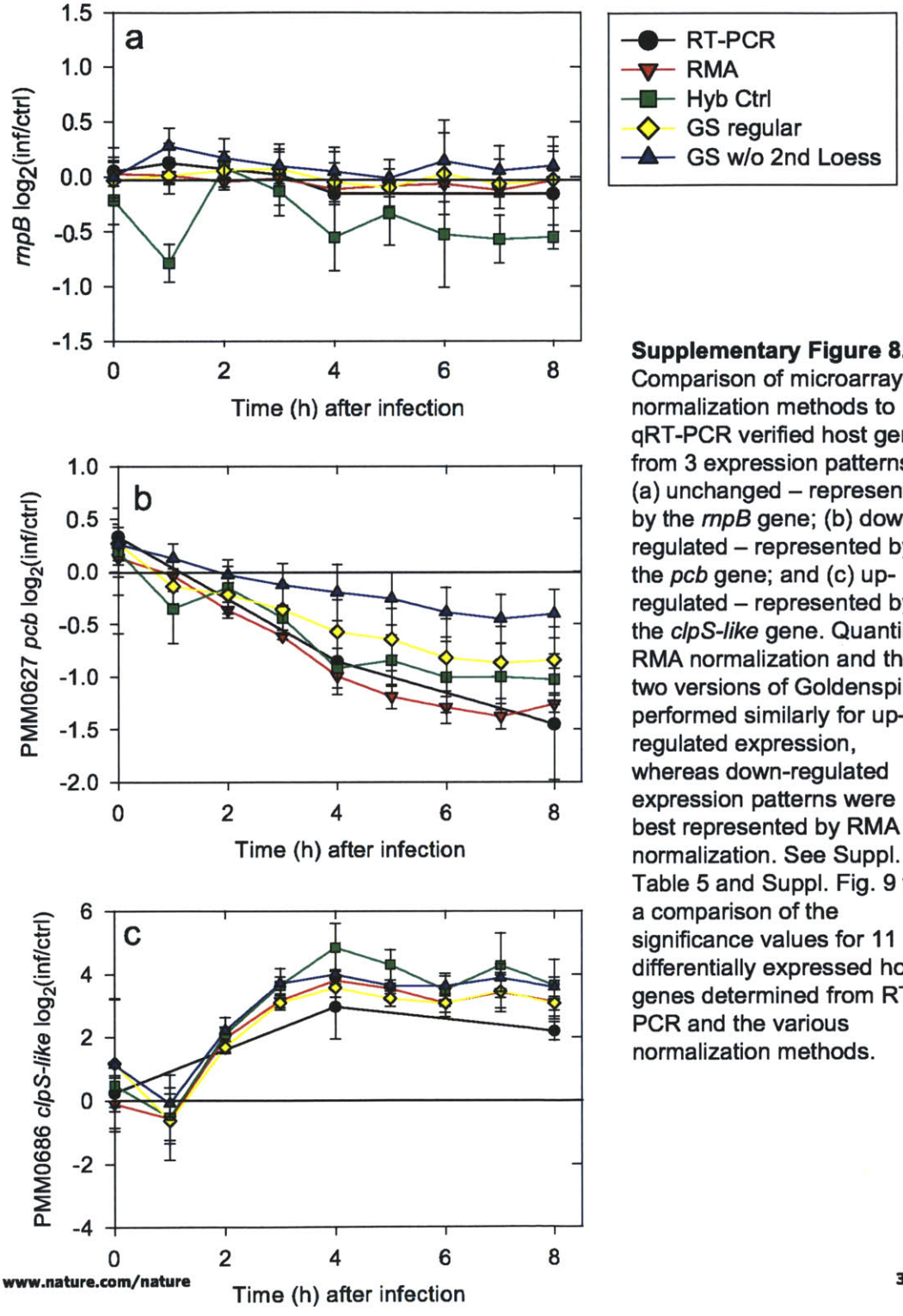
Supplementary Figure 5. Comparison of extracellular P-SSP7 quantification using a quantitative PCR (qPCR) assay for the phage DNA polymerase gene with (a) infective titer determined by the most probable number (MPN) assay and (b) total phage particles after staining with the DNA SYBR Green I stain and enumerated by epifluorescence microscopy. The linear regression for (a) is $y = 2.38x + 410324$, $R^2 = 0.87$; and (b) is $y = 0.84x - 132486$, $R2 = 0.97$. Note that qPCR quantification provides close to a 1:1 ratio with the SYBR stained particles, but was 2.5 fold higher than infective phage.



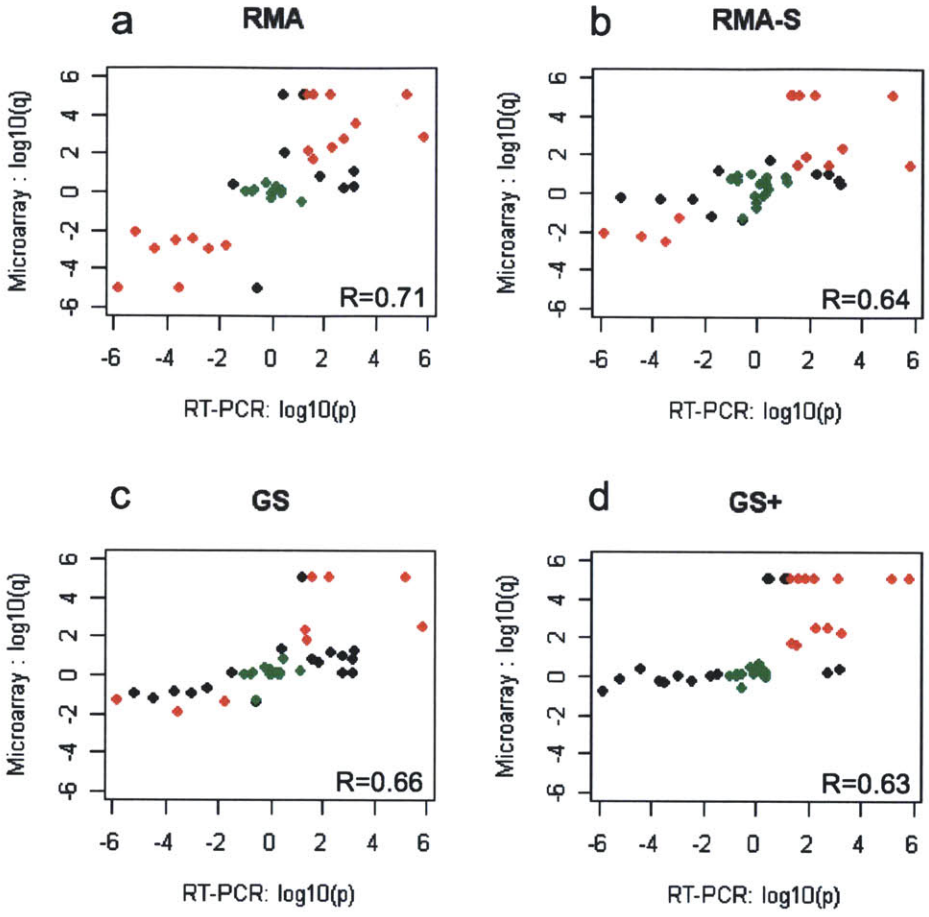
Supplementary Figure 6. qRT-PCR verification of phage gene expression patterns determined from microarray results. (a) Expression profiles for representative genes from each transcription cluster were analyzed by RT-PCR using gene specific primers. The results were normalized to *rnpB* (an internal control gene) to correct for potential differences in input RNA, and are presented relative to maximum levels for each gene. The results are shown on a logarithmic scale to better discern differences in expression patterns at the early time points which are low relative to maximal transcript levels. (b) Microarray results for the same representative genes shown to facilitate direct comparison to the RT-PCR results. Note that, as is commonly found, changes in expression determined by RT-PCR were orders of magnitude greater than by microarray analysis.



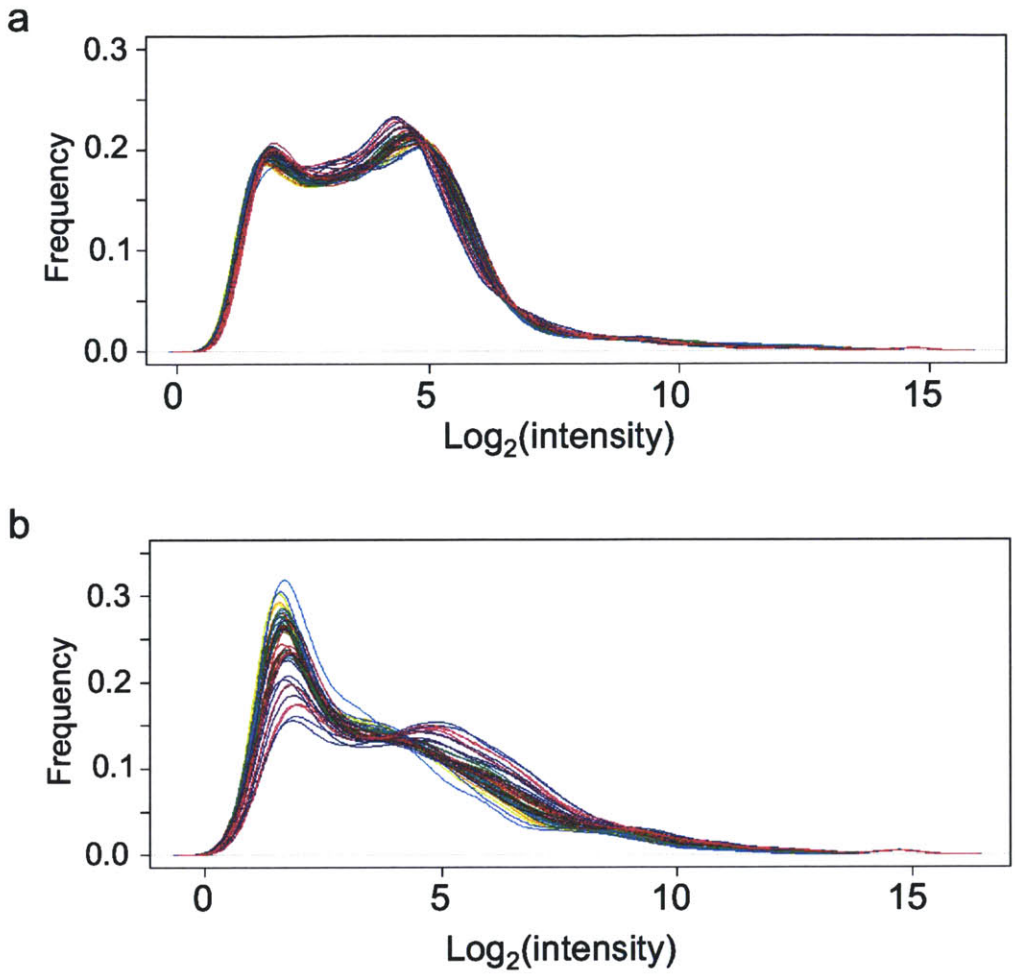
Supplementary Figure 7. qRT-PCR verification of host gene expression patterns determined from microarray analysis. Left panel (a, c, e) shows RT-PCR results and right panel (b, d, f) shows microarray results for representative genes displaying: (a, b) unchanged expression profile; (c, d) down-regulated genes; and (e, f) up-regulated genes. PMM1629 is the only gene whose up-regulated expression was not verified by RT-PCR (compare c and d at T=1 h). Suppl. Table 5 provides a direct comparison of the significance of differentially expressed from qRT-PCR and microarrays after quantile RMA normalization.



Supplementary Figure 8. Comparison of microarray normalization methods to qRT-PCR verified host genes from 3 expression patterns: (a) unchanged – represented by the *mpB* gene; (b) down-regulated – represented by the *pcb* gene; and (c) up-regulated – represented by the *clpS-like* gene. Quantile RMA normalization and the two versions of Goldenspike performed similarly for up-regulated expression, whereas down-regulated expression patterns were best represented by RMA normalization. See Suppl. Table 5 and Suppl. Fig. 9 for a comparison of the significance values for 11 differentially expressed host genes determined from RT-PCR and the various normalization methods.



Supplementary Figure 9. Comparison of the performance of different microarray normalization methods for detection of significant differences in gene expression as compared to RT-PCR analysis. (a) Quantile RMA normalization at the probe level; (b) RMA normalization based on spiked in hybridization controls; (c) Goldenspike (GS) without summary level normalization; and (d) Goldenspike with summary level normalization. R = Pearson correlation between RT-PCR and microarray analysis. Red and green symbols denote significant and insignificant differences respectively in gene expression called for both RT-PCR and microarray analysis, whereas black symbols denote discrepancies in significance calls between the RT-PCR and microarray analyses. The significance of differential expression for microarray analyses (q-values) was calculated using the Bayes t-test and significance of differential expression for RT-PCR was determined from a standard two-tailed t-test (p-values). Q-values <0.00001 were set to 0.00001 to ensure non-zero values. These findings show that quantile RMA normalization gave the highest correlation and the largest number of correctly identified differentially expressed genes, especially for downregulated genes.



Supplementary Figure 10. Density distribution of signal intensities for probe sets from each microarray after quantile RMA normalization: (a) All probe sets are displayed. The overall distribution for all arrays is similar due to the quantile normalization. (b) MED4 probe sets only are displayed. Note that different arrays displayed various distributions for the MED4 probe sets despite the similar distribution for all probe sets. Therefore quantile normalization can be used for this experiment without erasing differences in MED4 gene expression. Each line represents a different array.

Appendix B

Microarray normalization with Goldenspike

B.1 Introduction

In (Lindell et al., 2007, Appendix A), we reported on gene expression in *Prochlorococcus* MED4 during cyanophage infection. We considered a number of normalization methods and presented results using Robust Multi-Array Average (RMA). Here, I discuss alternate methods in more detail.

An Affymetrix GeneChip contains several short probes per feature (a feature is usually a gene). In the case of the *Prochlorococcus* MD4-9313 array, probes are also tiled across intergenic regions. Each probe is a perfect match (PM) to its target sequence, but each is also paired with a mismatch (MM) counterpart, which is close to but not the same as the probe sequence. Processing raw fluorescence values (from a .cel file) to produce expression values for each gene involves removing background signal and normalizing the signal level across the entire array. Thanks to the MM probes, an estimate of non-specific hybridization can be subtracted from the PM probes, hopefully yielding the "true" signal level of only the perfectly matching cDNA. Because there are several probes per feature, their values must be combined to yield one signal value for each feature.

The BioConductor *affy* package provides a variety of algorithms for each of these steps (Gautier et al., 2004). Choe et al. (2005) took advantage of a controlled spike-in dataset to test all possible combinations, finding the best option or options for each step. The *goldenspike* R package contains functions to invoke BioConductor using the these best practices. Because they found more than one option provided similar results in some steps, each among the best, *goldenspike* processes data

through eight separate pipelines, averages them, and uses a Bayesian t-test to assign false discovery rates (q-values). The *goldenspike* pipeline is summarized in fig. B-1.

Choe et al. (2005) also found that results were more accurate if, after combining separate probes into features, a second normalization was performed to remove biases that escaped the probe-level normalization. This proved to be problematic in the case of a phage infection experiment. Here, I present the results of our tests of *goldenspike*, with and without the second normalization.

B.2 Methods

B.2.1 RT-PCR

Experiments were performed on RNA samples from appendix A. RNA was reverse transcribed using gene-specific primers and SuperScript II (Invitrogen) with the following protocol:

- Heat RNA to 65° C, 5 minutes.
- Cool to 4° C.
- Add SuperScript enzyme, buffer, DTT, and Superase-In (Ambion).
- 42° C, 50 minutes.
- 70 ° C, 15 minutes.

QPCR was performed with a QuantiTect SYBR kit (Qiagen) and gene-specific primers with the following protocol:

- 95° C, 15 minutes.
- cycle:
- 95° C, 15 seconds.
- 56° C, 30 seconds.
- 72° C, 30 seconds.
- read.
- end cycle.

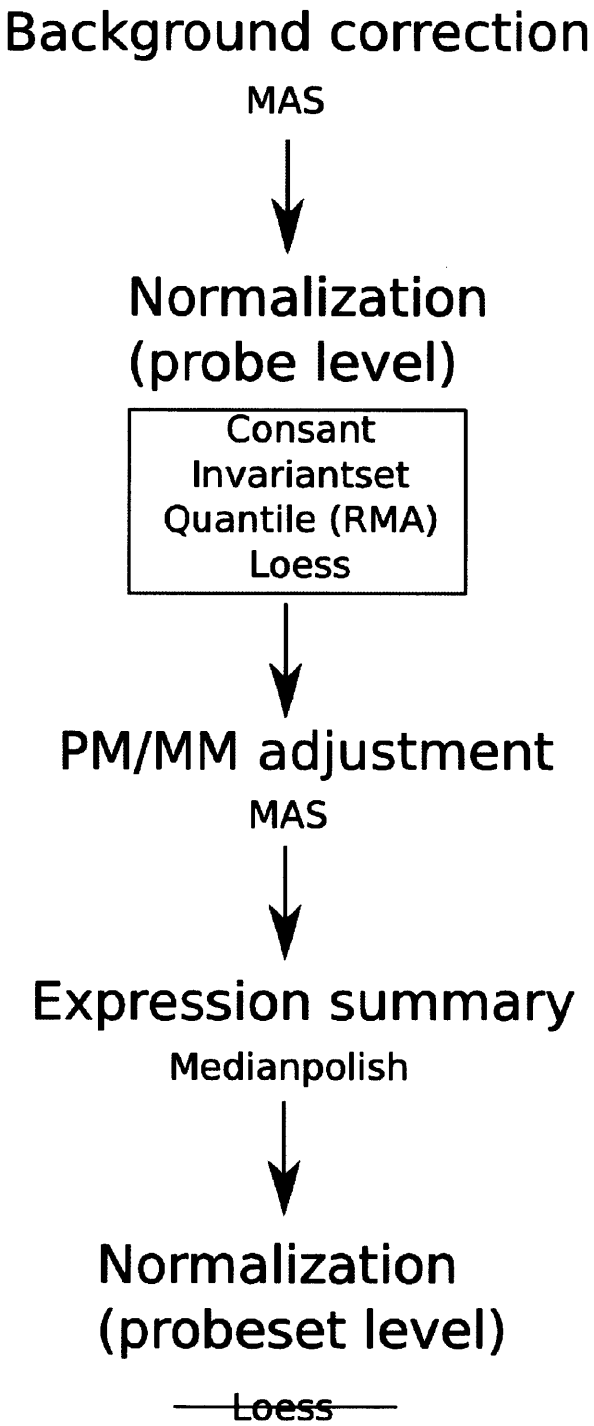


Figure B-1: The Goldenspike microarray analysis pipeline, adapted. The second normalization step, focus of this investigation, is highlighted. This figure is adapted from (Choe et al., 2005).

Locus	Gene	Direction	Primer
PMM_rnpB	rnpB – RNA of RNase P	F	TTGAGGAAAGTCCGGGCTC
		R	GCGGTATGTTTCTGTGGCACT
PMM0684	Unknown	F	CGCAAGGCAGCTTTTTAATC
		R	TCCATGTTTCAAACGCAGAG
PMM0686	clpS-like – protease adaptor	F	CAGTTGTAGATCCAAAGACAACG
		R	CAAGACAATTTGCTACGTGTTCA
PMM0819	Unknown	F	CCCAAGTGGTTGGCTTCTTA
		R	ATCCCAGGCTTTTTCCAAAT
PMM1284	phoH-like – phosphate stress induced ATPase	F	GTTTGTGCCGCCAGATTATT
		R	TGCTAATGGTGCGACTTCAA
PMM1501	rne – RNase E	F	AACCGCCTAGCACAGGATTA
		R	TGCTTTTTCGAGAGCGATTT

Table B.1: Primers used in the RT-PCR experiments. Reproduced from Appendix A.

- 72° C, 5 minutes.
- melt curve 50 - 90° C

All expression changes were normalized against *rnpB*. Six genes were tested: *rnpB*, PMM0684, PMM0686, PMM0819, PMM1284 and PMM1501 (Table B.1).

B.2.2 Array normalization

I examined Affymetrix array data from appendix A. I invoked *goldenspike*, specifically the make.expr.summaries function, to generate feature-level signal values from each of the *goldenspike* pipelines, and do.paired.comparisons to perform the statistical test on the whole set. I also modified make.expr.summaries to save all data just before the second normalization, and I used a modified version of do.paired.comparisons to read that data back and proceed normally.

B.3 Results

Using the default pipeline, at the 4-hour timepoint, expression of 26 non-RNA MED4 genes was found to be significantly changed. Of these, 14 were upregulated ($q < 0.05$). However, without the second normalization, 45 were found to be significantly changed. Only 4 of these were significantly upregulated.

We compared selected upregulated genes to our RT-PCR results and found that without the normalization, array results were closer to RT-PCR (fig. B-2).

B.4 Conclusion

The second normalization compensates for system-wide deviations in observed expression. However, phage infection is responsible for a true system-wide decrease in expression (Ueno and Yonesaki, 2004), and to "compensate" for it may be an error.

However, there is the possibility that the smaller declines in expression of certain genes is of biological significance. That decline may reflect a greater "push" by the cell to raise expression against the "pull" of overall RNA degradation.

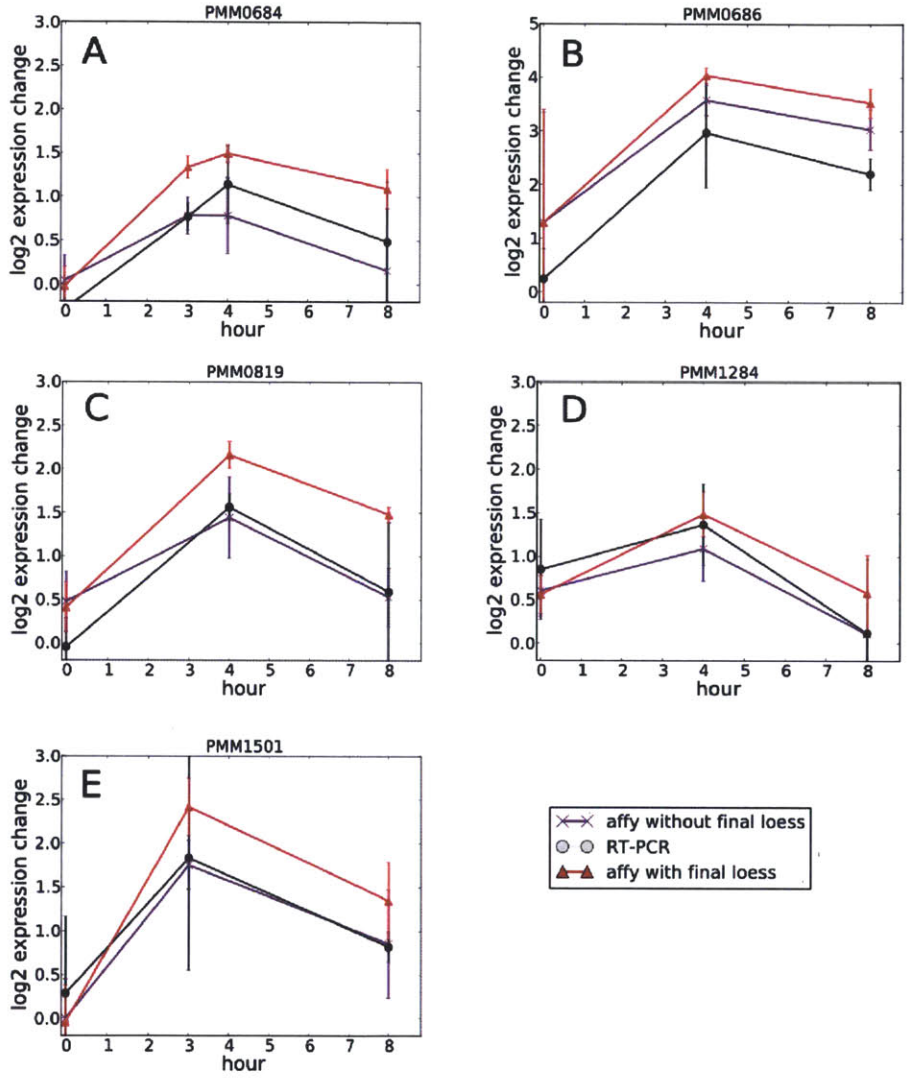


Figure B-2: Timeseries of 5 selected genes as measured by RT-PCR, microarray with second Loess, and microarray without second Loess. Reported as $\log_2$(infected/control). RT-PCR results are normalized to *rnpB*. (A) PMM0684 (B) PMM0686 (C) PMM0819 (D) PMM1284 (E) PMM1501.

Appendix C

Temporal dynamics of *Prochlorococcus* ecotypes in the Atlantic and Pacific oceans

Author contributions

R.R.M., A.C.M., J.F.L., and E.R.Z. designed ITS primers. R.R.M. and A.C. carried out QPCR measurements. G.C.K. performed light shock experiments. R.R.M. and S.W.C. wrote the manuscript.

The ISME Journal (2010) 4, 1252–1264
© 2010 International Society for Microbial Ecology All rights reserved 1751-7362/10
www.nature.com/ismej

ORIGINAL ARTICLE

Temporal dynamics of *Prochlorococcus* ecotypes in the Atlantic and Pacific oceans

Rex R Malmstrom[1], Allison Coe[1], Gregory C Kettler[2], Adam C Martiny[1,3], Jorge Frias-Lopez[1,4], Erik R Zinser[1,5] and Sallie W Chisholm[1]

[1]*Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA;* [2]*Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA;* [3]*Departments of Earth System Science and Ecology and Evolutionary Biology, University of California, Irvine, CA, USA;* [4]*The Forsyth Institute, Boston, MA, USA and* [5]*Department of Microbiology, University of Tennessee, Knoxville, TN, USA*

To better understand the temporal and spatial dynamics of *Prochlorococcus* populations, and how these populations co-vary with the physical environment, we followed monthly changes in the abundance of five ecotypes—two high-light adapted and three low-light adapted—over a 5-year period in coordination with the Bermuda Atlantic Time Series (BATS) and Hawaii Ocean Time-series (HOT) programs. Ecotype abundance displayed weak seasonal fluctuations at HOT and strong seasonal fluctuations at BATS. Furthermore, stable 'layered' depth distributions, where different *Prochlorococcus* ecotypes reached maximum abundance at different depths, were maintained consistently for 5 years at HOT. Layered distributions were also observed at BATS, although winter deep mixing events disrupted these patterns each year and produced large variations in ecotype abundance. Interestingly, the layered ecotype distributions were regularly reestablished each year after deep mixing subsided at BATS. In addition, *Prochlorococcus* ecotypes each responded differently to the strong seasonal changes in light, temperature and mixing at BATS, resulting in a reproducible annual succession of ecotype blooms. Patterns of ecotype abundance, in combination with physiological assays of cultured isolates, confirmed that the low-light adapted eNATL could be distinguished from other low-light adapted ecotypes based on its ability to withstand temporary exposure to high-intensity light, a characteristic stress of the surface mixed layer. Finally, total *Prochlorococcus* and *Synechococcus* dynamics were compared with similar time series data collected a decade earlier at each location. The two data sets were remarkably similar—testimony to the resilience of these complex dynamic systems on decadal time scales.

The ISME Journal (2010) **4**, 1252–1264; doi:10.1038/ismej.2010.60; published online 13 May 2010
Subject Category: Microbial population and community ecology
Keywords: *Prochlorococcus*; *Synechococcus*; ecotype; time-series; HOT; BATS

Introduction

A key challenge facing marine microbiology is to understand how microbial diversity and biogeochemical cycles are linked, and to eventually incorporate this understanding into conceptual and predictive ocean models. Physiological and genetic analyses of cultured isolates, as well as metagenomic studies of whole communities (Venter *et al.*, 2004; DeLong *et al.*, 2006), are uncovering more and more about the metabolic potential of microbes comprising these assemblages. In addition, time series studies are revealing how the composition of microbial communities varies over both time and space (Fuhrman *et al.*, 2006; Carlson *et al.*, 2009; Treusch *et al.*, 2009). Coupling the spatial and temporal dynamics of specific microbial groups with insights into their metabolic potential is essential for developing a quantitative understanding of the roles these microbes have in marine ecosystems.

Over the past 20 years, investigations of the unicellular cyanobacterium *Prochlorococcus* have provided insight into both the biogeography and metabolic potential of this group. *Prochlorococcus* is typically the most abundant photoautotroph in tropical and subtropical waters (Campbell *et al.*, 1994; Partensky *et al.*, 1999), and its abundance varies seasonally at some locations (Campbell *et al.*, 1997; DuRand *et al.*, 2001). Studies of cultured isolates have revealed the optimal light and temperature levels differ among the strains (Moore *et al.*, 1998, 2002; Moore and Chisholm, 1999; Zinser *et al.*, 2007), as do the nutrient pools available to them (Moore *et al.*, 2002, 2005).

Correspondence: S Chisholm, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, 15 Vassar St, Cambridge, MA, 02139, USA.
E-mail: chisholm@mit.edu
Received 4 January 2010; revised 22 March 2010; accepted 24 March 2010; published online 13 May 2010

Genomic analyses of isolates (Rocap *et al.*, 2003; Coleman *et al.*, 2006; Kettler *et al.*, 2007), and metagenomic analyses of natural communities (DeLong *et al.*, 2006; Martiny *et al.*, 2006, 2009a), have provided additional insights into the genetic diversity, metabolic potential and evolutionary history of the group. The ability to examine the abundance and distribution of *Prochlorococcus* in the wild, and measure its genomic and metabolic properties in the lab and field, make *Prochlorococcus* a model system for advancing our understanding the ecology of marine microbes.

Prochlorococcus is composed of several clades that are physiologically and phylogenetically distinct with respect to their optimal light and temperature environments (Moore *et al.*, 1998; Moore and Chisholm, 1999; Rocap *et al.*, 2002; Johnson *et al.*, 2006; Kettler *et al.*, 2007; Zinser *et al.*, 2007). These clades have been referred to as 'ecotypes' (Moore *et al.*, 1998; Moore and Chisholm, 1999; Rocap *et al.*, 2002) following the broader historical designation for genetically distinct subgroups within a species that are adapted to specific environments (Turesson, 1922; Clausen *et al.*, 1940). They do not necessarily conform to the more recent and narrowly defined 'ecotype' concept developed by Cohan and others (Cohan, 2001; Cohan and Perry, 2007), which has become a notable model for exploring the theoretical basis for divergence among bacteria (Ward *et al.*, 2006; Frasier *et al.*, 2009). A more detailed discussion of the different uses of the term 'ecotype' is provided by Coleman and Chisholm, (2007).

The abundance of *Prochlorococcus* ecotypes in various oceanic regions has been studied extensively (West and Scanlan, 1999; Ahlgren *et al.*, 2006; Bouman *et al.*, 2006; Johnson *et al.*, 2006; Zinser *et al.*, 2006, 2007). Members of the two high-light adapted (HL) ecotypes, eMIT9312 and eMED4, are most abundant in the upper regions of the euphotic zone, whereas low-light adapted (LL) ecotypes such as eNATL and eMIT9313 are most abundant in the lower euphotic zone (West *et al.*, 2001; Johnson *et al.*, 2006; Zinser *et al.*, 2007). These distribution patterns agree well with differences in the optimal light levels for representative ecotype strains (Moore and Chisholm, 1999; Zinser *et al.*, 2007). Furthermore, eMED4 tends to dominate in cooler, higher latitude waters, whereas eMIT9312 dominates warmer, lower latitude waters (Johnson *et al.*, 2006; Zwirglmaier *et al.*, 2007); also in good agreement with temperature optima of representative strains. Water column stability and nutrient concentrations have also been correlated with ecotype abundance (Bouman *et al.*, 2006; Johnson *et al.*, 2006), and community structure (Martiny *et al.*, 2009b), although the underlying causalities of these relationships are not well understood.

The environmental factors influencing *Prochlorococcus* ecotype abundance, such as light, temperature and water column mixing, vary over time and can display seasonal patterns, thus we might expect ecotype dynamics to do the same. Although we have analyzed ecotype variability over a span of a few days (Zinser *et al.*, 2007), extensive time-series studies have not been conducted, and thus little is known about the dynamics of *Prochlorococcus* ecotypes on the scale of months to years. Analyses on these time scales should help refine our understanding of ecotype/environment interactions, and provide data sets for testing models designed to explore the dynamics of phytoplankton community structure (Follows *et al.*, 2007).

To this end, we followed the spatial and temporal dynamics of *Prochlorococcus* ecotypes at monthly intervals over 5 years in coordination with the Bermuda Atlantic Time Series (BATS) and the Hawaii Ocean Time-series (HOT) programs (Karl and Lukas, 1996; Steinberg *et al.*, 2001). Both programs are focused on oligotrophic, open ocean sites where *Prochlorococcus* is found in abundance (Campbell *et al.*, 1994; DuRand *et al.*, 2001). However, the physics and chemistry of these locations differ, most notably by the stronger seasonal mixing events at BATS and the higher inorganic phosphate concentrations at HOT (Wu *et al.*, 2000; Cavender-Bares *et al.*, 2001; Steinberg *et al.*, 2001). Here, we explore how changes in environmental factors such as mixing, light and temperature are related to ecotype abundance and distribution, as well as how they co-vary temporally over 5 years. We also examine some of the emergent patterns in the field data through studies of light-shock tolerance in cultured isolates of different *Prochlorococcus* ecotypes.

Materials and methods

Sample collection

Beginning in November 2002, flow cytometry and qPCR samples were collected over a 5-year period during monthly cruises for the Bermuda Atlantic Time Series (BATS) and Hawaii Ocean Time-series (HOT) programs. Additional samples were collected bi-weekly between February and April at BATS. Samples were collected from 12 depths (1, 10, 20, 40, 60, 80, 100, 120, 140, 160, 180 and 200 m) at the BATS site (~5 nautical mile radius around 31° 40′N, 64° 10′ W), and from 12 depths (5, 25, 45, 60, 75, 85, 100, 115, 125, 150, 175 and 200 m) at Sta. ALOHA (~5 nautical mile radius around 22° 45′N, 158° 00′W). These locations are referred to BATS and HOT throughout the article for simplicity.

Flow cytometry

Whole seawater samples were immediately fixed with glutaraldehyde (final conc. 0.125% v/v) for 10 min, frozen in liquid nitrogen, and stored at −80 °C until samples could be processed in the laboratory using an influx flow cytometer

(Becton Dickinson, Franklin Lakes, NJ, USA). *Prochlorococcus* and *Synechococcus* populations were identified and quantified based on their unique autofluorescence and scatter signals (Olson *et al.*, 1990a, 1990b). *Prochlorococcus* could not always be clearly distinguished in the upper 40 m at BATS from June–September, thus these profiles were excluded from depth-integrated counts in Figure 2. Flow cytometry profiles from early 2003 were not processed.

Primer re-design for ecotype eSS120

New primers for the eSS120 ecotype were designed in ARB (Ludwig *et al.*, 2004), using a large database of environmental ITS sequences (Martiny *et al.*, 2009b). These primers, 5′-AACAAACTTTCTCCTG GGCT-3′ and 5′-AGTTGATCAGTGGAGGTAAG-3′, matched 87% of known eSS120 ITS sequences, but did not target MIT9211. The specificity of the primers was initially tested against cultured isolates from other ecotypes such as MIT9313, MIT9312, MED4 and NATL2a. These tests confirmed specificity within the dynamic range of the assay (~ 5 to 5×10^5 cells ml^{-1}). Specificity was also confirmed by cloning and sequencing ITS regions amplified from field samples from BATS and HOT using the new primers. All amplified ITS sequences clustered with strains SS120 and MIT9211, both members of the eSS120 ecotype (Kettler *et al.*, 2007), in a bootstrapped ($n = 100$) neighbor-joining tree constructed with the Bosque software package (Ramirez-Flandes and Ulloa, 2008) (Supplementary Figure 1). The new primers substantially increased counts of the SS120 ecotype (Supplementary Materials).

Quantitative PCR

Samples for qPCR were collected and processed as described previously (Ahlgren *et al.*, 2006; Zinser *et al.*, 2006, 2007) with two small modifications. First, reaction volumes were reduced from 25 µl to 15 µl and performed in 384-well plates on the Light Cycler 480 for samples collected after 2003. Second, concentrations of eMIT9313-specific primers were increased to 5 µM to improve sensitivity. Estimated abundances that fell below the lowest value of the standard curve were set to the theoretical detection limit of 0.65 cells ml^{-1}. Samples were excluded if their melt curves contained multiple peaks or peaks different from those in the DNA standards to ensure only the targeted ecotypes were quantified. Missing data were determined by linear interpolation when abundance estimates were available for the depths immediately above and below the missing value.

Environmental data extraction

Temperature, salinity and potential density were downloaded from the HOT and BATS websites. Missing data were determined by linear interpolation when values were available for the depths immediately above and below the missing data point. Mixed layer depths, light attenuation coefficients and *Prochlorococcus* and *Synechococcus* abundance (1991–1995) were also downloaded directly from the HOT website. Mixed layer depths and attenuation coefficients at BATS were calculated from bottle-derived profiles and SeaWiFS Profiling Multichannel Radiometer profiles of photosynthetically active radiation (PAR) collected by Bermuda Bio-Optics Program. The mixed layer depth was determined when potential density differed from surface values by >0.125 kg m^{-3}. Attenuation coefficients at BATS were calculated by linear regression of log-transformed PAR values; only profiles with an $R^2 > 0.993$ were used.

Solar irradiance

Surface irradiance was determined from SeaWiFS-derived estimates of daily-integrated PAR. Eight-day means of daily integrated PAR values from March 2000 to July 2006 were calculated from a 27 km by 27 km region around BATS and HOT (White *et al.*, 2007). A two component Fourier model with a period of 1 year was fitted to PAR data, producing an R^2 of 0.90 and 0.93 at BATS and HOT, respectively. Modeled PAR data were used to estimate surface irradiance on the day of sample collection. This model was necessary to account for seasonal variability in solar flux due to changes in day length and solar azimuth. These changes result in a roughly two-fold difference in daily-integrated solar flux between summer and winter (for example, $\sim 20\,000$–$56\,000$ mE m^{-2} d^{-1} at BATS, and $\sim 31\,000$–$58\,000$ mE m^{-2} d^{-1} at HOT, using this model).

The relationships between ecotype abundance and PAR plotted in Figure 1 and Supplementary Figure 2 were determined using robust locally weighted linear regression (LOWESS) in MATLAB. Robust LOWESS is more resistant to outliers (Cleveland, 1979), which are defined in MATLAB's robust LOWESS function as data outside six mean absolute deviations.

Time series and other statistical analyses

Integrated ecotype abundance, surface PAR, and mixed layer depth were log-transformed, detrended, and resampled at a regular monthly interval to meet the mathematical requirements for time series analyses (Legendre and Legendre, 1998). Detrended data were calculated as the residuals of linear regression of log-transformed data against time. Detrended data were resampled every 30.44 days, which is equivalent to 12 measurements per year, using linear interpolation. Coefficients of autocorrelation and cross-correlation were determined in MATLAB, and s.d. calculated as $n^{-1/2}$, where n is the number or resampled data points. Spectral

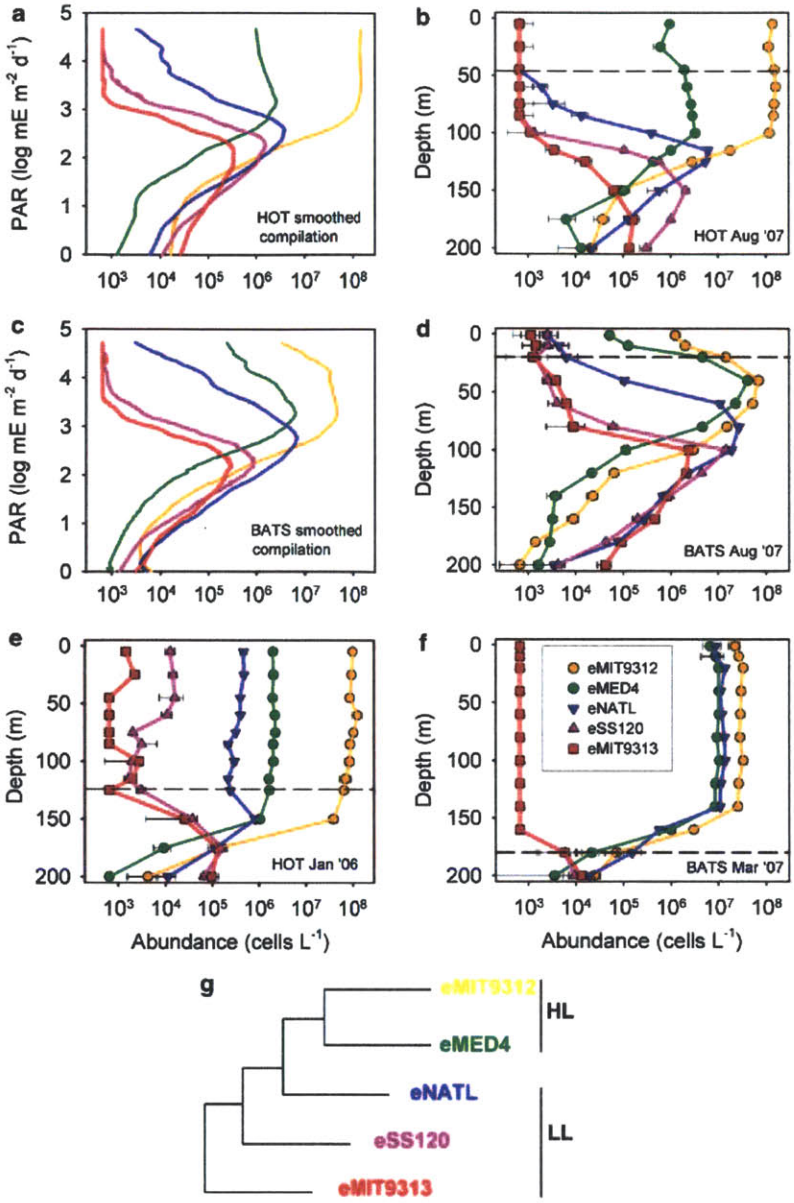


Figure 1 Ecotype distribution patterns in relation to depth and photosynthetically active radiation (PAR). A smoothed illustration of the relationship between ecotype abundance and irradiance, calculated from a compilation of data from all 5 years, was plotted for HOT (**a**) and BATS (**c**). Data were smoothed by locally weighted regression of ecotype abundance against irradiance. Smoothed data exclude profiles with a mixed layer depth >100 m. Representative depth profiles of ecotype abundance at HOT (**b**, **e**) and BATS (**d**, **f**) during periods of stratification and deep mixing. Dashed lines mark the mixed layer depth, and error bars represent one s.d. An illustration of the color scheme and general phylogenetic relationships among ecotypes after (Rocap *et al.*, 2002; Kettler *et al.*, 2007) is displayed in (**g**). This tree serves only as reference to the other panels; branch lengths are not to scale.

analysis by discrete Fourier transformation calculated using the Fast Fourier Transformation (FFT) algorithm, was also performed in MATLAB. The power spectral density was estimated as the absolute value of FFT^2.

Differences among ecotypes in the average depth of maximum abundance were tested using repeated-measures ANOVA, followed by a Tukey *post-test* ($\alpha = 0.05$), using MATLAB.

Non-parametric partial correlation coefficients (Spearman R) were calculated in MATLAB to determine the relationship between abundance and temperature while controlling for the influence of light.

Light shock experiments
Serial batch cultures of axenic *Prochlorococcus* strains MED4, NATL2a and SS120 were grown in Sargasso seawater-based Pro99 media (Moore *et al.*, 2007) and illuminated by cool white fluorescent lamps. Cultures were transferred at least four times to acclimatize them to 35 $\mu E\,m^{-2}\,s^{-1}$ of continuous light. Duplicate acclimated cultures that were in log-phase growth were then exposed to 400 $\mu E\,m^{-2}\,s^{-1}$ for 4 h before being returned to their initial light levels. *In vivo* chlorophyll fluorescence was measured before and after light shock with a 10-AU fluorometer (340–500 nm excitation and 680 nm emission filters), and cell counts determined by flow cytometry as described above.

Results and Discussion

Ecotype distribution with light and temperature
Striking similarities between HOT and BATS emerged when data from all depths and all 5 years were combined for a synoptic analysis of *Prochlorococcus* ecotype abundance along irradiance/depth gradients (Figures 1a and b; Supplementary Figure 2). As expected, the two high-light adapted ecotypes, eMIT9312 and eMED4, were most abundant at higher irradiances, with abundance dropping off sharply below 1500 $mE\,m^{-2}\,d^{-1}$ of photosynthetically active radiation (Figures 1a and b; Supplementary Figures 2a and e). In contrast, two low-light adapted ecotypes, eSS120 and eMIT9313, usually reached maximum abundances between 100 and 250 $mE\,m^{-2}\,d^{-1}$, and were typically at or near detection limits at irradiances $>1500\,mE\,m^{-2}\,d^{-1}$. The eNATL group, also low-light adapted had an intermediate distribution, reaching maximum abundance at 300–600 $mE\,m^{-2}\,d^{-1}$. Unlike the other low light ecotypes, eNATL abundance was occasionally high at irradiance levels $>1500\,mE\,m^{-2}\,d^{-1}$ (Figures 1a and b; Supplementary Figures 2c and h), which is consistent with hypothesis that eNATL can tolerate exposure to higher light levels than eSS120 and eMIT9313 (Coleman and Chisholm, 2007; Zinser *et al.*, 2007).

This consistent relationship between irradiance and ecotype abundance results in a 'layered" depth distribution at both locations (Figures 1a–d). At HOT, for example, HL ecotype eMIT9312 tended to reach maximum abundance at shallower depths than did its fellow HL ecotype eMED4 (Table 1). LL ecotypes also partitioned the water column at HOT, with eNATL abundance peaking at significantly shallower depths than eSS120 and eMIT9313 (Table 1). Similar patterns in ecotype distribution were also observed at BATS (Table 1; Figures 1b and d), except for during deep mixing events, defined here as when the mixed layer depth was >100 m. During periods of deep mixing physical homogenization appears to overwhelm biological partitioning of the water column, resulting in uniform depth distributions (Figures 1e and f). Therefore, data from periods of deep mixing were removed from statistical analysis of depth distributions.

Table 1 Average depth of maximum ecotype abundance at HOT and BATS (mean ± STD)

Ecotype	*HOT (m)*	*BATS (m)*
eMIT9312	42 ± 27^{A}	54 ± 28^{A}
eMED4	71 ± 29^{B}	60 ± 26^{A}
eNATL	105 ± 18^{C}	87 ± 22^{B}
eSS120	118 ± 26^{D}	101 ± 28^{C}
eMIT9313	128 ± 30^{D}	107 ± 25^{C}

Samples collected when the mixed layer depth was >100 m were excluded. Values with different superscripts are significantly different (Tukey post test of repeated measures ANOVA; $\alpha = 0.05$).

It is remarkable that, with the exception of periods of deep mixing, the general patterns of ecotype abundance appear relatively consistent in the Atlantic and Pacific despite substantial differences in the chemical and physical environment. Furthermore, these patterns are consistent with those found previously in a variety of ocean regions (West and Scanlan, 1999; West *et al.*, 2001; Bouman *et al.*, 2006; Johnson *et al.*, 2006; Zinser *et al.*, 2006, 2007). That is, the ecotypes tend to partition the water column by depth, with eNATL reaching maximum abundance in between the peaks of HL ecotypes MED4 and MIT9312, and other LL ecotypes. The similarities in the distributions along depth/light gradients suggest that *Prochlorococcus* ecotypes are responding in a consistent fashion to irradiance regardless of geography.

Temporal dynamics: depth-integrated Prochlorococcus *and* Synechococcus *populations*
Using flow cytometry, we measured the abundance of *Prochlorococcus*, and its close relative *Synechococcus* (Rocap *et al.*, 2002), to provide an overall framework for exploring *Prochlorococcus* ecotype dynamics. At BATS, the depth-integrated *Prochlorococcus* population displayed a strong seasonal pattern, reaching the highest levels in the late summer and fall, and the lowest in the late winter during the annual deep mixing events (Figure 2a). *Synechococcus* displayed the inverse pattern, and even occasionally exceeded the abundance of *Prochlorococcus* during deep mixing events (Figure 2b). This is the same pattern reported by DuRand *et al*, (2001) for *Prochlorococcus* and *Synechococcus* at BATS from 1990 to 1994 (Figures 2a and b). The concordance between these two data sets is remarkable, given that they are separated by more than a decade.

Variations in *Prochlorococcus* and *Synechococcus* abundance were much less dramatic at HOT. Integrated abundance of *Prochlorococcus* varied just over 2-fold throughout the time series, and did not always reach maximal abundance in the summer or minimal abundance in the winter (Figure 2c).

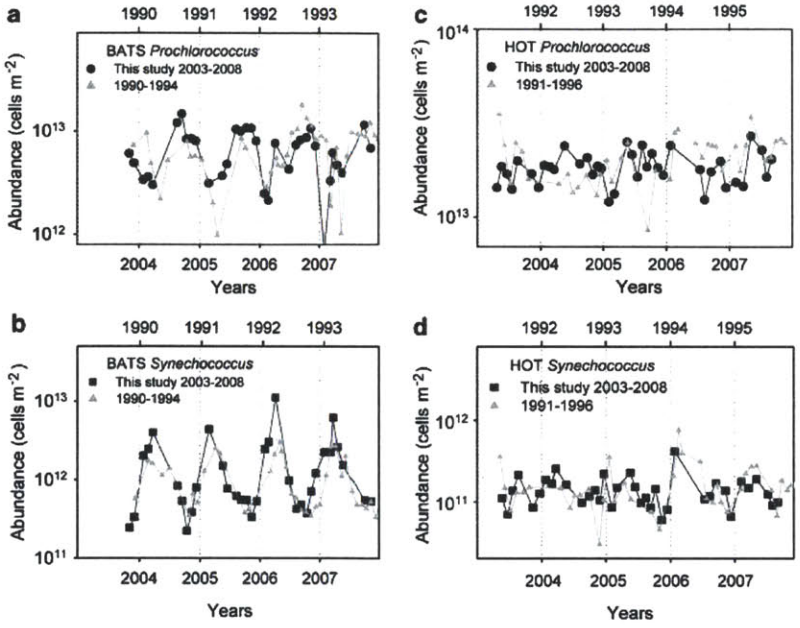


Figure 2 Abundance of *Prochlorococcus* and *Synechococcus* at BATS and HOT determined by flow cytometry (bottom *x* axis). Abundance levels from this study are compared with those collected at BATS from 1990 to 1994 (DuRand *et al.*, 2001) (**a**, **b**), and at HOT from 1991 to 1995 (**c**, **d**) (top *x* axis), which were also determined by flow cytometry.

Synechococcus still displayed an annual pattern, tending to reach peak abundance during winter, but they were never as abundant as *Prochlorococcus*, in contrast to what was observed at BATS. These abundance and variability levels are also consistent with those observed over a decade ago at HOT (Figures 2c and d)

Temporal dynamics: depth-integrated ecotype abundance patterns

As was seen in the total *Prochlorococcus* population, the depth-integrated (0–200 m) abundance of all five ecotypes followed clear annual patterns at BATS (Figure 3a). Spectral analysis of each ecotype revealed dominant peaks in the power spectrum at a period of 1 year (Supplementay Figure 3), and autocorrelations displayed a sinusoidal pattern, with peaks in autocorrelation every 12 months (Supplementary Figure 4). Furthermore, fitting a single component Fourier series with a period of 1 year to each ecotype produced R^2 values ranging from 0.48 (eMIT9313) to 0.67 (eSS120), indicating that most of the variability in integrated abundance could be accounted for by an annual oscillation.

Annual variations in temperature and mixed layer depth were smaller at HOT than at BATS, as were the variations in ecotype abundance (Figure 3b). Although spectral analysis did reveal that all ecotypes had a peak in the power spectrum at a period of 1 year, this was not the only strong signal,

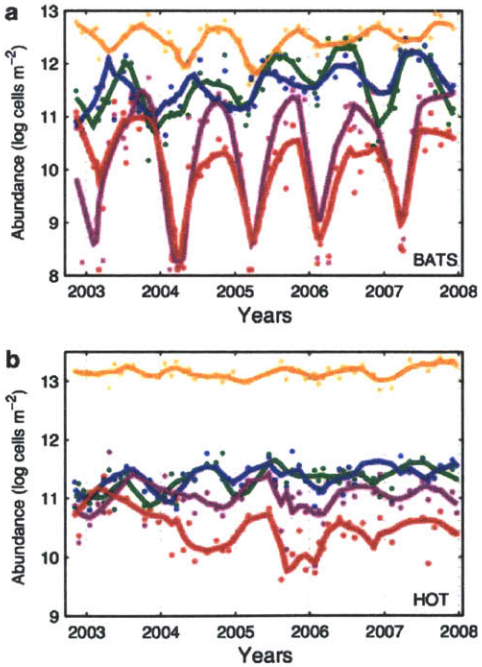


Figure 3 Integrated ecotype abundance (0–200 m) at BATS (**a**) and HOT (**b**) from 2003 to 2008. Solid lines represent abundances smoothed by locally weighted regression for eMIT9312 (yellow), eMED4 (green), eNATL (blue), eSS120 (purple) and eMIT9313 (red).

particularly for the LL ecotypes (Supplementary Figure 5). Fitting a single component Fourier series with a period of 1 year to each ecotype produced R^2 values ranging from only 0.17 (eMIT9312) to 0.33 (eMIT9313). In addition, autocorrelations did not display the same strong sinusoidal pattern as seen at BATS (compare Supplementary Figure 3 with Supplementary Figure 6). Thus, while there was a component of annual variability to ecotype abundance at HOT, there was also a variability at periods greater and less than 1 year. This suggests that when annual environmental variations are more moderate, the impact of intra- and inter-annual events, such as passage of mesoscale eddies or El Niño/La Niña oscillations—both known to influence primary production and phytoplankton community composition at HOT (Karl *et al.*, 1995; Letelier *et al.*, 2000; Corno *et al.*, 2007; Bibby *et al.*, 2008)—may become more pronounced.

Although ecotype abundance followed a clear annual cycle at BATS, the cycles were not synchronized among ecotypes, that is, different ecotypes reached peak abundance at different times. In each of the 5 years, the eNATL ecotype reached its maximum integrated abundance about 4 months after winter deep mixing events (Table 2), typically in June (Supplementary Figure 7). Abundance of eMED4 peaked roughly 1 month later, whereas the LL-clades eSS120 and eMIT9313 reached maximal abundance about 8 months after the deep mixing event (Table 2). HL ecotype eMIT9312 also reached maximal abundances at around the same time as eSS120 and eMIT9313, typically in October or November (Supplementary Figure 7). The regularity of this pattern indicates that each ecotype responded to changes in environmental conditions in different, yet consistent ways, resulting in an annually repeating succession of ecotypes.

The succession of *Prochlorococcus* ecotypes is in some ways reminiscent of classical phytoplankton succession models, although it is occurring at a much smaller phylogenetic scale; all *Prochlorococcus* differ in 16S rRNA sequence by <3% (Moore *et al.*, 1998), which would collectively constitute a single bacterial species by conventional standards (Stackebrandt and Goebel, 1994). Interestingly, two recent time series studies at BATS have also uncovered annual cycles and succession patterns in other microbial groups, most notably the SAR11 clade (Carlson *et al.*, 2009; Treusch *et al.*, 2009). For example, one SAR11 subgroup reaches peak abundance in surface waters during the summer, whereas another subgroup peaks in the winter (Carlson *et al.*, 2009). SAR11 bacteria are the most abundant heterotrophs at BATS and are major consumers of dissolve organic compounds (Morris *et al.*, 2002; Malmstrom *et al.*, 2005), whereas *Prochlorococcus* is the most abundant photoautotroph and a substantial source of dissolved organics (DuRand *et al.*, 2001; Bertilsson *et al.*, 2005). In addition, SAR11 bacteria and *Prochlorococcus* are both major consumers of small compounds like amino acids and dimethylsulfoniopropionate (DMSP) (Zubkov *et al.*, 2003; Malmstrom *et al.*, 2004; Vila-Costa *et al.*, 2006; Michelou *et al.*, 2007), which are significant sources of C, N and S to marine microbial communities. Therefore, it seems plausible that succession in these two abundant groups could be linked through the production of, and competition for, dissolved organic compounds. Uncovering potential links in the dynamics of the dominant microbial groups presents a future challenge.

Table 2 Correlation and cross-correlation between integrated ecotype abundance (0–200 m) and mixed layer depth at BATS

Ecotype	*Correlation coeff. (lag in months)*	*Max. cross-correlation coeff. (lag in months)*
eMIT9312	−0.06 (0)	0.61* (8)
eMED4	−0.62* (0)	0.58* (5)
eNATL	−0.54* (0)	0.65* (4)
eSS120	−0.60* (0)	0.66* (8)
eMIT9313	−0.46* (0)	0.61* (8)

The coefficient of cross-correlation represents the relationship between abundance at 1 month and the mixed layer depth from previous months. The maximum positive cross-correlation is reported alongwith the delay in months (lag) between abundance and mixed layer depth when correlation is greatest. *Indicates $P<0.05$.

Ecotype abundance in different regions of the euphotic zone

Variations in light and temperature levels, which impact the growth of *Prochlorococcus* (Moore and Chisholm, 1999; Johnson *et al.*, 2006; Zinser *et al.*, 2007), are greater in surface waters than at depth. Thus integrating abundance across the entire water column likely obscures important features in ecotype dynamics. To get a more detailed understanding of the temporal and spatial dynamics of the ecotypes, we analyzed integrated abundance in three sections of the euphotic zone (0–60 m, 60–120 m and 120–200 m).

The HL-adapted ecotype eMIT9312 displayed similar abundance patterns in the upper (0–60 m) and middle (60–120 m) euphotic zone at BATS, but below 120 m the seasonal cycle was out of phase with surface cycles by several months (Figure 4a; Figure 5a). Abundance in the lower euphotic was positively correlated with mixed layer depth (Spearman $R=0.5$; $P<0.05$), and abundance peaks occurred simultaneously with annual deep mixing events (Figure 4a). This suggests that it is the transport of surface populations below 120 m, and not *in situ* growth that may be responsible for most of the annual variations in abundance of eMIT9312 in the lower euphotic zone at BATS.

As with eMIT9312, patterns of eMED4 abundance in the upper and middle euphotic zone also differed from those in the lower zone at BATS (Figure 5b), with strong spikes in abundance below 120 m

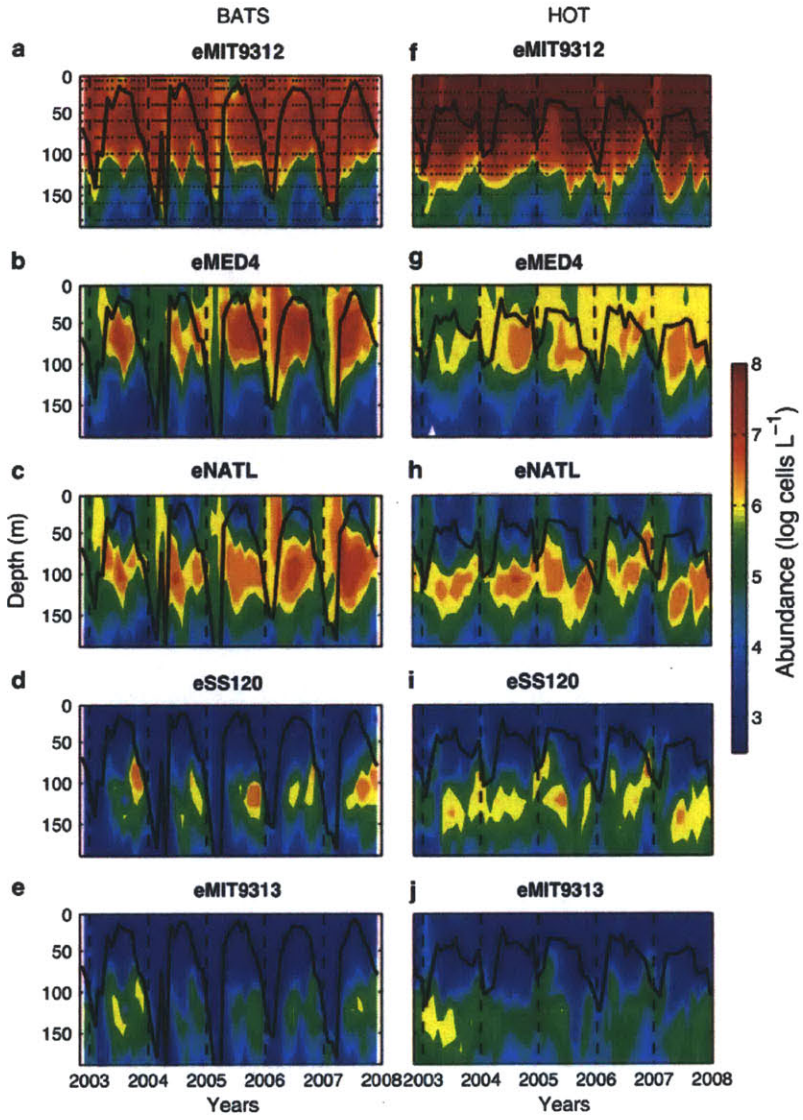


Figure 4 Ecotype abundance at BATS (**a–e**) and HOT (**f–j**) from 2003 to 2008. Sampling depths at HOT and BATS are indicated by the points overlaying (**a** and **f**). The solid lines indicate the mixed layer depth.

accompanying deep mixing (Figure 4b). But while the eMED4 and eMIT9312 dynamics were synchronized below 120 m, these HL-adapted ecotypes were out of synchronization by several months in the upper 120 m (Figures 5a and b). That is, each year at BATS, eMED4 typically reached peak abundance in July–August, whereas eMIT9312 reached maximum abundance in October–November (Supplementary Figure 7). At HOT, in contrast, these two HL ecotypes did not display these offset repeating patterns.

We hypothesize that different temperature sensitivities of eMED4 and eMIT9312 explain, at least in part, the differences in their temporal dynamics at BATS. That is, strains belonging to the eMED4 clade have lower temperature optima than those belonging to the eMIT9312 clade (Johnson *et al.*, 2006; Zinser *et al.*, 2007). If this differential trait is universal among cells belonging to the eMED4 and eMIT9313 clades, then this would allow eMED4 cells to accumulate during the first half of the year when temperatures were low, whereas the higher temperature optimum of eMIT9312 would limit their accumulation until later in the season when temperatures were high—as was observed at BATS. In fact, the abundance of eMED4 in the upper 60 m was negatively correlated with temperature when light levels were taken into account (partial correlation coeff. –0.43; $P<0.05$), whereas eMIT9312 abundance was positively correlated with temperature

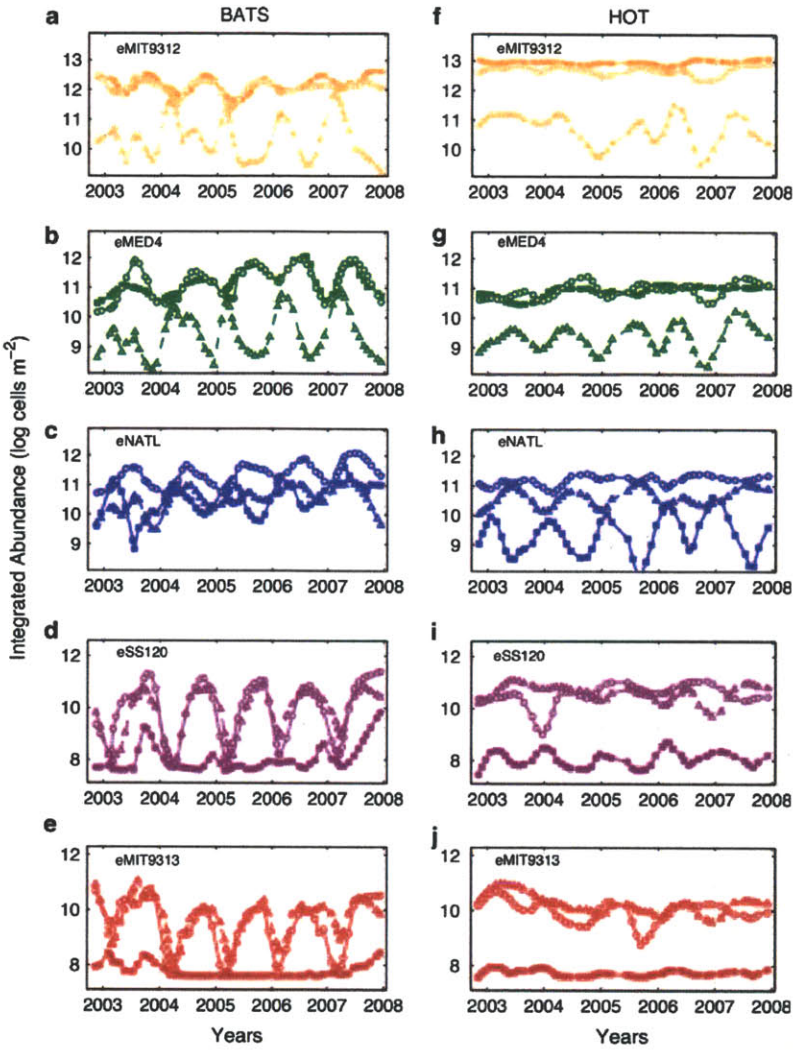


Figure 5 Integrated ecotype abundance in different regions of the euphotic zone at BATS (**a–e**) and HOT (**f–j**). The regions from 0–60 m, 60–120 m and 120 m are identified by filled squares, open circles and open triangles, respectively. Abundance was smoothed using locally weighted regression to eliminate low frequency variation.

(partial correlation coeff. 0.30; $P<0.05$). The potential influence of temperature on the temporal distribution of HL ecotypes at BATS is analogous to its inferred influence on their geographic distribution: eMED4 dominates in cooler, higher latitude waters, and eMIT9312 dominates warmer, lower latitude waters (Johnson *et al.*, 2006; Zwirglmaier *et al.*, 2007).

Abundance patterns of eNATL also differed among regions of the euphotic zone, and the similarity of these patterns between BATS and HOT was striking. At both locations, abundance in the upper 60 m typically peaked 3–4 months earlier than at 60–120 m and 120–200 m (Figures 4c and h; Figures 5c and h). Integrated abundance in the upper 60 m reached maximum levels during deep mixing events, and decreased as the mixed layer depth shoaled (Figures 4c and h). This suggests that annual peaks in abundance in the upper 60 m were due, at least in part, to vertical transport of deeper cells to surface waters via mixing. While it remains unclear if members of the LL-adapted eNATL clade were able to grow at high light levels found in the upper euphotic zone, the net accumulation eNATL cells throughout the water column during periods of deep mixing confirmed that the eNATL clade was able to at least tolerate temporary exposure to high irradiance.

In contrast to eNATL, the abundance of the other two LL-adapted ecotypes, eSS120 and eMIT9313, did not increase dramatically in the upper 60 m during periods of deep mixing. In fact, their

residue that is known to dramatically reduce activity in homologs (Doi *et al.*, 1992; Goosen and Moolenaar, 2008), suggesting a diminished ability to repair UV-damaged DNA in eSS120 and eMIT9313 ecotypes. Indeed, the high-light adapted strain MED4, which encodes photolyase, has a greater tolerance to UV exposure than low-light adapted strain MIT9313 (Osburne *et al.*, 2010), which lacks photolyase but encodes pyrimidine dimmer glycosylase. Thus, the protection from UV exposure provided by photolyase may explain, at least in part, why the eNATL clade can better survive transport to UV-rich surface waters.

Conclusions

Clear patterns in the temporal and spatial distribution of *Prochlorococcus* ecotypes emerge from this study. For example, ecotype abundance follows a strong annual pattern at BATS, whereas ecotype abundance has only a weak annual pattern at HOT. In addition, ecotypes at BATS follow an annual succession pattern, but a similar pattern is not observed at HOT. These patterns are consistent with what we have learned from physiological assays on cultured isolates with regards to the light optima, temperature optima, and light-shock tolerance of different ecotypes. That is, the distinct distribution patterns at both HOT and BATS can be explained, at least in general terms, by the consistent and predictable responses of ecotypes to changes in the light, temperature and mixing at each location.

Analyses of inorganic nutrient concentrations and ecotype abundance were not possible as nutrient levels were below detection in the upper 100 m at BATS throughout most of this study. However, while nutrients undoubtedly influence growth rates and standing stocks of *Prochlorococcus*, their influence on specific ecotypes may not be apparent, as evidenced by the fact that general patterns can be explained without them. Indeed, recent work suggests that nitrate concentrations might impact the composition of *Prochlorococcus* populations at finer phylogenetic levels than the ecotypes examined in our study (Martiny *et al.*, 2009b). Exploring how, and at which temporal, spatial and phylogenetic scales, the chemical and biological environment influence *Prochlorococcus* abundance and diversity presents a future challenge.

Results from this study help set the stage for coupling patterns in temporal and spatial dynamics of ecotypes with insights into their metabolic potential. The next step is to understand how *Prochlorococcus* ecotypes, or even sub-groups within these ecotypes, might differ in terms of nutrient usage, dissolved organic matter production and consumption, and other metabolic processes. This understanding will come from additional studies involving strain isolation, metagenomic comparisons, and large-scale single-cell genomics.

Acknowledgements

We thank Michael Lomas and the BATS team for sample collection at Bermuda, and David Karl, Matthew Church, and the HOT team for sample collection at Hawaii. We also thank Angel White and Ricardo Letelier for assistance with deriving solar flux data and attenuation coefficients. Norm Nelson and David Court generously provided light data from the Bermuda Bio Optics Program. We thank Daniele Veneziano for advice on statistical analysis. This work was funded by grants from the National Science Foundation, NSF STC Center for Microbial Oceanography: Research and Education, and the Gordon and Betty Moore Foundation.

References

Ahlgren NA, Rocap G, Chisholm SW. (2006). Measurement of *Prochlorococcus* ecotypes using real-time polymerase chain reaction reveals different abundances of genotypes with similar light physiologies. *Environ Microbiol* **8**: 441–454.

Bertilsson S, Berglund O, Pullin MJ, Chisholm SW. (2005). Release of dissolved organic matter by *Prochlorococcus*. *Vie Et Milieu-Life and Environment* **55**: 225–231.

Bibby TS, Gorbunov MY, Wyman KW, Falkowski PG. (2008). Photosynthetic community responses to upwelling in mesoscale eddies in the subtropical north atlantic and pacific oceans. *Deep-Sea Res Part II-Top Stud Oceanogr* **55**: 1310–1320.

Bouman HA, Ulloa O, Scanlan DJ, Zwirglmaier K, Li WK, Platt T *et al.* (2006). Oceanographic basis of the global surface distribution of *Prochlorococcus* ecotypes. *Science* **312**: 918–921.

Campbell L, Liu HB, Nolla HA, Vaulot D. (1997). Annual variability of phytoplankton and bacteria in the subtropical north pacific ocean at station ALOHA during the 1991–1994 ENSO event. *Deep-Sea Res Part I-Oceanogr Res Pap* **44**: 167.

Campbell L, Nolla HA, Vaulot D. (1994). The importance of *prochlorococcus* to community structure in the central north pacific-ocean. *Limnol Oceanogr* **39**: 954–961.

Carlson CA, Morris R, Parsons R, Treusch AH, Giovannoni SJ, Vergin K. (2009). Seasonal dynamics of SAR11 populations in the euphotic and mesopelagic zones of the northwestern Sargasso Sea. *ISME J* **3**: 283–295.

Cavender-Bares KK, Karl DM, Chisholm SW. (2001). Nutrient gradients in the western north atlantic ocean: relationship to microbial community structure and comparison to patterns in the pacific ocean. *Deep-Sea Res Part I-Oceanogr Res Pap* **48**: 2373–2395.

Clausen J, Keck DD, Hiesey WM. (1940). Experimental studies on the nature of species. I. Effects of varied environments on western North American plants. *Carnegie Institute of Washington* **520**.

Cleveland WS. (1979). Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc* **74**: 829–836.

Coleman ML, Chisholm SW. (2007). Code and context: *Prochlorococcus* as a model for cross-scale biology. *Trends Microbiol* **15**: 398–407.

Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, Delong EF *et al.* (2006). Genomic islands

and the ecology and evolution of *Prochlorococcus*. *Science* **311**: 1768–1770.
Cohan FM. (2001). Bacterial species and speciation. *Systematic Biol* **50**: 513–524.
Cohan FM, Perry EB. (2007). A systematics for discovering the fundamental units of bacterial diversity. *Curr Biol* **17**: R373–R386.
Corno G, Karl DM, Church MJ, Letelier RM, Lukas R, Bidigare RR *et al.* (2007). Impact of climate forcing on ecosystem processes in the North Pacific Subtropical Gyre. *J Geophys Res* **112**: C04021, doi:10.1029/2006JC003730.
Doi T, Recktenwald A, Karaki Y, Kikuchi M, Morikawa K, Ikehara M *et al.* (1992). The role of the basic amino acid cluster and Glu-23 in pyrimidine dimer glycosylase activity of T4-endonuclease-V. *Proc Natl Acad Sci USA* **89**: 9420–9424.
DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU *et al.* (2006). Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**: 496–503.
DuRand MD, Olson RJ, Chisholm SW. (2001). Phytoplankton population dynamics at the bermuda atlantic time-series station in the sargasso sea. *Deep-Sea Res Part II-Top Stud Oceanogr* **48**: 1983–2003.
Frasier C, Alm EJ, Polz MF, Spratt BG, Hanage WP. (2009). The bacterial species challenge: making sense of genetic and ecological diversity. *Science* **323**: 741–746.
Follows MJ, Dutkiewicz S, Grant S, Chisholm SW. (2007). Emergent biogeography of microbial communities in a model ocean. *Science* **315**: 1843–1846.
Fuhrman JA, Hewson I, Schwalbach MS, Steele JA, Brown MV, Naeem S. (2006). Annually reoccurring bacterial communities are predictable from ocean conditions. *Proc Natl Acad Sci USA* **103**: 13104–13109.
Goosen N, Moolenaar GF. (2008). Repair of UV damage in bacteria. *DNA Repair* **7**: 353–379.
Havaux M, Guedeney G, He QF, Grossman AR. (2003). Elimination of high-light-inducible polypeptides related to eukaryotic chlorophyll a/b-binding proteins results in aberrant photoacclimation in Synechocystis PCC6803. *Biochimica Et Biophysica Acta-Bioenergetics* **1557**: 21–33.
He QF, Dolganov N, Bjorkman O, Grossman AR. (2001). The high light-inducible polypeptides in Synechocystis PCC6803 - Expression and function in high light. *J Biol Chem* **276**: 306–314.
Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EMS, Chisholm SW. (2006). Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science* **311**: 1737–1740.
Karl DM, Letelier R, Hebel D, Tupas L, Dore J, Christian J *et al.* (1995). Ecosystem changes in the north pacific subtropical gyre attributed to the 1991–92 El-Nino. *Nature* **373**: 230–234.
Karl DM, Lukas R. (1996). The hawaii ocean time-series (HOT) program: background, rationale and field implementation. *Deep-Sea Res Part II-Top Stud Oceanogr* **43**: 129–156.
Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S *et al.* (2007). Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *Plos Genetics* **3**: 2515–2528.
Legendre P, Legendre L (eds). (1998). Ecological data series. In: *Numerical Ecology: Second English Addition*. Elsevier Science BV: Amsterdam, pp 637–691.
Letelier RM, Karl DM, Abbott MR, Flament P, Freilich M, Lukas R *et al.* (2000). Role of late winter mesoscale events in the biogeochemical variability of the upper water column of the north pacific subtropical gyre. *J Geophysical Res-Oceans* **105**: 28723–28739.
Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar Buchner A *et al.* (2004). ARB: a software environment for sequence data. *Nucleic Acids Res* **32**: 1363–1371.
Malmstrom RR, Cottrell MT, Elifantz H, Kirchman DL. (2005). Biomass production and dissolved organic matter assimilation by SAR11 bacteria in the northwest atlantic ocean. *Appl Environ Microbiol* **71**: 2979–2986.
Malmstrom RR, Kiene RP, Cottrell MT, Kirchman DL. (2004). Contribution of SAR11 bacteria to dissolved dimethylsulfoniopropionate and amino acid uptake in the North Atlantic Ocean. *Appl Environ Microbiol* **70**: 4129–4135.
Martiny AC, Coleman ML, Chisholm SW. (2006). Phosphate acquisition genes in *Prochlorococcus* ecotypes: Evidence for genome-wide adaptation. *Proc Natl Acad Sci USA* **103**: 12552–12557.
Martiny AC, Kathuria S, Berube PM. (2009a). Widespread metabolic potential for nitrite and nitrate assimilation among *Prochlorococcus* ecotypes. *Proc Natl Acad Sci USA* **106**: 10787–10792.
Martiny AC, Tai APK, Veneziano D, Primeau F, Chisholm SW. (2009b). Taxonomic resolution, ecotypes and the biogeography of *Prochlorococcus*. *Environ Microbiol* **11**: 823–832.
Michelou VK, Cottrell MT, Kirchman DL. (2007). Light-stimulated bacterial production and amino acid assimilation by cyanobacteria and other microbes in the north atlantic ocean. *Appl Environ Microbiol* **73**: 5539–5546.
Moore LR, Chisholm SW. (1999). Photophysiology of the marine cyanobacterium *Prochlorococcus*: ecotypic differences among cultured isolates. *Limnol Oceanogr* **44**: 628–638.
Moore LR, Coe A, Zinser ER, Saito MA, Sullivan MB, Lindell D *et al.* (2007). Culturing the marine cyanobacterium Prochlorococcus. *Limnol Oceanog* **5**: 353–362.
Moore LR, Ostrowski M, Scanlan DJ, Feren K, Sweetsir T. (2005). Ecotypic variation in phosphorus acquisition mechanisms within marine picocyanobacteria. *Aquat Microb Ecol* **39**: 257–269.
Moore LR, Post AF, Rocap G, Chisholm SW. (2002). Utilization of different nitrogen sources by the marine cyanobacteria *Prochlorococcus* and *Synechococcus*. *Limnol Oceanogr* **47**: 989–996.
Moore LR, Rocap G, Chisholm SW. (1998). Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature* **393**: 464–467.
Morris RM, Rappé MS, Connon SA, Vergin KL, Siebold WA, Carlson CA *et al.* (2002). SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420**: 806–810.
Olson RJ, Chisholm SW, Zettler ER, Altabet MA, Dusenberry JA. (1990a). Spatial and temporal distributions of prochlorophyte picoplankton in the north-atlantic ocean. *Deep-Sea Res Part a-Oceanographic Res Papers* **37**: 1033–1051.
Olson RJ, Chisholm SW, Zettler ER, Armbrust EV. (1990b). Pigments, size, and distribution of synechococcus

in the north-atlantic and pacific oceans. *Limnol Oceanogr* **35**: 45–58.

Osburne MS, Holmbeck BM, Frias-Lopez J, Steen R, Huang K, Kelly L *et al.* (2010). UV hyper-resistance in *Prochlorococcus* MED4 results from a single base pair deletion just upstream of an operon encoding nudix hydrolase and photolyase. *Environ Microbiol* Published Online: 23 March 2010; DOI: 10.1111/j.1462-2920.2010.02203.x.

Partensky F, Garczarek L. (2010). Prochlorococcus: advantages and limits of minimalism. *Annu Rev Marin Sci* Vol. 2 305–331.

Partensky F, Hess WR, Vaulot D. (1999). Prochlorococcus:a marine photosynthetic prokaryote of global significance. *Microbiol Mol Biol Rev* **63**: 106.

Promnares K, Komenda J, Bumba L, Nebesarova J, Vacha F, Tichy M. (2006). Cyanobacterial small chlorophyll-binding protein ScpD (HliB) is located on the periphery of photosystem II in the vicinity of PsbH and CP47 subunits. *J Biol Chem* **281**: 32705–32713.

Ramirez-Flandes S, Ulloa O. (2008). Bosque: integrated phylogenetic analysis software. *Bioinformatics* **24**: 2539–2541.

Rocap G, Distel DL, Waterbury JB, Chisholm SW. (2002). Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. *Appl Environ Microbiol* **68**: 1180–1191.

Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA *et al.* (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424**: 1042–1047.

Sancar GB. (2000). Enzymatic photoreactivation: 50 years and counting. *Mutation Res-Fundamental Mol Mechanisms Mutagenesis* **451**: 25–37.

Six C, Finkel ZV, Irwin AJ, Campbell DA. (2009). Light variability illuminates niche-partitioning among marine picocyanobacteria. *PLoS ONE* **2**: e1341.

Stackebrandt E, Goebel BM. (1994). Taxonomic note: a place for DNA:DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int J Syst Bacteriol* **44**: 846–849.

Steinberg DK, Carlson CA, Bates NR, Johnson RJ, Michaels AF, Knap AH. (2001). Overview of the US JGOFS Bermuda atlantic time-series study (BATS): a decade-scale look at ocean biology and biogeochemistry. *Deep-Sea Res Part II-Top Stud Oceanogr* **48**: 1405–1447.

Treusch AH, Vergin KL, Finlay LA, Donatz MG, Burton RM, Carlson CA *et al.* (2009). Seasonality and vertical structure of microbial communities in an ocean gyre. *ISME J* **3**: 1148–1163.

Turesson G. (1922). Species and the variety as ecological units. *Hereditas* **3**: 100–113.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA *et al.* (2004). Environmental genome shotgun sequencing of the sargasso sea. *Science* **304**: 66–74.

Vila-Costa M, Simo R, Harada H, Gasol JM, Slezak D, Kiene RP. (2006). Dimethylsulfoniopropionate uptake by marine phytoplankton. *Science* **314**: 652–654.

Ward DM, Bateson MM, Ferris MJ, Kühl M, Wieland A, Koeppel A *et al.* (2006). Cyanobacterial ecotypes in the microbial mat community of mushroom spring (yellowstone national park, wyoming) as species-like units linking microbial community composition, structure and function. *Philos Trans R Soc London (Biol)* **361**: 1997–2008.

West NJ, Scanlan DJ. (1999). Niche-partitioning of *Prochlorococcus* populations in a stratified water column in the eastern North Atlantic Ocean. *Appl Environ Microbiol* **65**: 2585–2591.

West NJ, Schönhuber WA, Fuller NJ, Amann RI, Rippka R, Post AF *et al.* (2001). Closely related *Prochlorococcus* genotypes show remarkably different depth distributions in two oceanic regions as revealed by *in situ* hybridization using 16S rRNA-targeted oligonucleotides. *Microbiology* **147**: 1731–1744.

White AE, Spitz YH, Letelier RM. (2007). What factors are driving summer phytoplankton blooms in the North Pacific Subtropical Gyre? *J Geophys Res* **112**: C12006, doi:10.1029/2007JC004129.

Wu JF, Sunda W, Boyle EA, Karl DM. (2000). Phosphate depletion in the western north atlantic ocean. *Science* **289**: 759–762.

Yao D, Kieselbach T, Komenda J, Promnares K, Prieto MA, Tichy M *et al.* (2007). Localization of the small CAB-like proteins in photosystem II. *J Biol Chem* **282**: 267–276.

Zinser ER, Coe A, Johnson ZI, Martiny AC, Fuller NJ, Scanlan DJ *et al.* (2006). Prochlorococcus ecotype abundances in the North Atlantic Ocean as revealed by an improved quantitative PCR method. *Appl Environ Microbiol* **72**: 723–732.

Zinser ER, Johnson ZI, Coe A, Karaca E, Veneziano D, Chisholm SW. (2007). Influence of light and temperature on *Prochlorococcus* ecotype distributions in the Atlantic Ocean. *Limnol Oceanogr* **52**: 2205–2220.

Zubkov MV, Fuchs BM, Tarran GA, Burkill PH, Amann R. (2003). High rate of uptake of organic nitrogen compounds by *Prochlorococcus* cyanobacteria as a key to their dominance in oligotrophic oceanic waters. *Appl Environ Microbiol* **69**: 1299–1304.

Zwirglmaier K, Heywood JL, Chamberlain K, Woodward EMS, Zubkov MV, Scanlan DJ. (2007). Basin-scale distribution patters of picocyanobacterial lineages in the Atlantic Ocean. *Environ Microbiol* **9**: 1278–1290.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)

Supplemental data for Appendix C

New estimates of ecotype eSS120 abundance

In previous studies, qPCR results could account for most *Prochlorococcus*, as determined by flow cytometry, but discrepancies between flow cytometry and qPCR did exist (Ahlgren et al. 2006; Zinser et al. 2006; Zinser et al. 2007). These discrepancies were greatest at deeper depths, suggesting that some low-light-adapted (LL) cells were being missed. When the ecotype-specific primers were evaluated using a new dataset of environmental ITS sequences (Martiny et al. 2009), it was clear that the existing primers for LL-adapted eSS120 clade were not suitable. In fact, the primers only recognized the SS120 type strain, thus explaining why counts of eSS120 were typically at or below detection limit in previous studies (Zinser et al. 2006; Zinser et al. 2007). Primers for eSS120 were re-designed, and these primers matched 87% of eSS120 ITS sequences, although they did not match MIT9211.

The new primers dramatically increased qPCR counts of the eSS120 clade (Table C.1). In eleven BATS profiles where data were available from both old and new primers, the integrated abundance (0-200m) of eSS120 increased from 36-fold to as much as 1,837-fold (Table C.1). However, estimates of eSS120 were not substantially different during periods of deep mixing at BATS, defined here when the mixed layer depth was >95m, since abundance was near or below detection limits with both primers sets. Integrated abundance of eSS120 also jumped by 20 to 269-fold in the fourteen HOT profiles where comparisons were possible. These new estimates, in contrast to previous studies, indicate that eSS120 reaches abundances comparable to those the LL ecotype eNATL and HL ecotype eMED4 (Fig. C-3, Fig. C-4).

The differences between the total flow cytometry counts and total qPCR counts decreased with new estimates of the eSS120 clade. When integrated across the upper 200m, the qPCR results from the five ecotypes accounted for 85 +/- 21% of total *Prochlorococcus* at HOT and 84% +/- 39% at BATS (mean +/- SD). However, discrepancies between qPCR and flow cytometric data were still greatest below 100m. For example, total qPCR counts equaled 41 +/- 22% of total *Prochlorococcus* at HOT when integrated from 100-200m. At this point, our current quantitative tools do not yet capture the entire diversity of *Prochlorococcus*, but still enable us to follow the vast majority of *Prochlorococcus* with just five ecotype primer sets.

Location	Cruise	Date	Mixed Layer Depth (m)	Old Primers (cells/m^2)	New Primers (cells/m^2)	Fold Difference
BATS	171	12/11/02	82	3.60E+008	1.80E+008	0.5
BATS	172	02/05/03	141	1.30E+008	6.60E+008	5
BATS	173	02/21/03	98	1.40E+008	1.10E+009	8
BATS	173.5	03/04/03	99	1.30E+008	1.30E+008	1
BATS	174	03/23/03	102	2.70E+008	1.90E+008	1
BATS	175	04/22/03	32	5.10E+008	3.30E+010	64
BATS	176	05/20/03	44	5.80E+008	1.10E+011	192
BATS	177	07/04/03	17	7.60E+008	3.90E+010	52
BATS	178	07/15/03	20	7.10E+008	2.60E+010	36
BATS	179	08/12/03	24	1.60E+009	2.70E+011	164
BATS	180	09/19/03	27	1.50E+009	2.90E+011	192
BATS	181	10/07/03	51	9.40E+008	3.10E+011	330
BATS	182	11/04/03	75	3.00E+008	5.50E+011	1837
BATS	183	12/02/03	84	4.80E+008	1.40E+011	289
BATS	184	01/27/04	151	1.30E+008	4.50E+009	35
BATS	184.5	02/14/04	169	1.30E+008	1.30E+008	1
HOT	141	11/04/02	67	6.70E+008	7.60E+010	114
HOT	142	11/25/02	83	5.90E+008	1.80E+010	30
HOT	143	12/19/02	82	1.70E+009	5.20E+010	30
HOT	144	01/18/03	117	1.20E+009	3.50E+010	29
HOT	146	03/29/03	43	2.70E+009	5.80E+010	22
HOT	147	04/24/03	51	2.70E+009	2.20E+011	81
HOT	148	05/21/03	44	1.20E+009	1.20E+011	103
HOT	149	06/20/03	47	4.20E+009	1.70E+011	40
HOT	151	08/21/03	57	9.70E+008	1.80E+011	190
HOT	152	10/15/03	67	1.70E+009	3.50E+010	20
HOT	153	11/10/03	69	1.00E+009	1.70E+011	166
HOT	154	12/21/03	50	6.60E+008	1.30E+011	201
HOT	155	01/22/04	104	1.40E+009	2.40E+011	177
HOT	157	03/20/04	95	6.80E+008	1.80E+011	269

Table C.1: Comparison qPCR primers for eSS120 ecotype. Integrated abundances (0-200m) were calculated with the original and redesigned primers at several time points at both HOT and BATS.

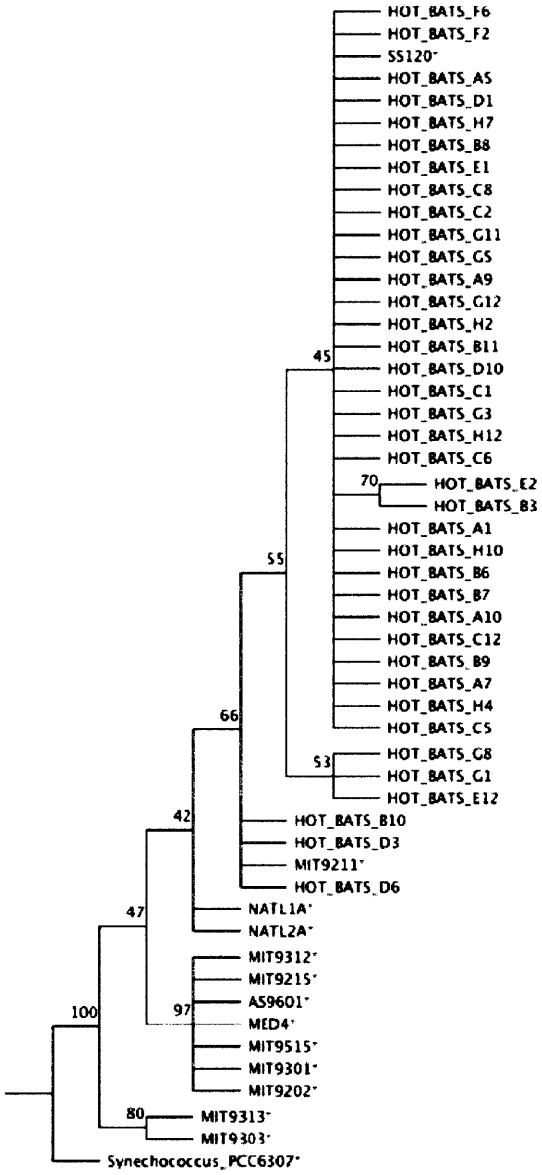


Figure C-1: Neighbor-joining tree of ITS sequences amplified from both HOT and BATS using re-designed qPCR primers specific for eSS120 ecotype. Amplified ITS sequences are identified by name "HOT_BATS." Previously sequenced genomes are marked with an *. Bootstrap values >40 are indicated (n=100).

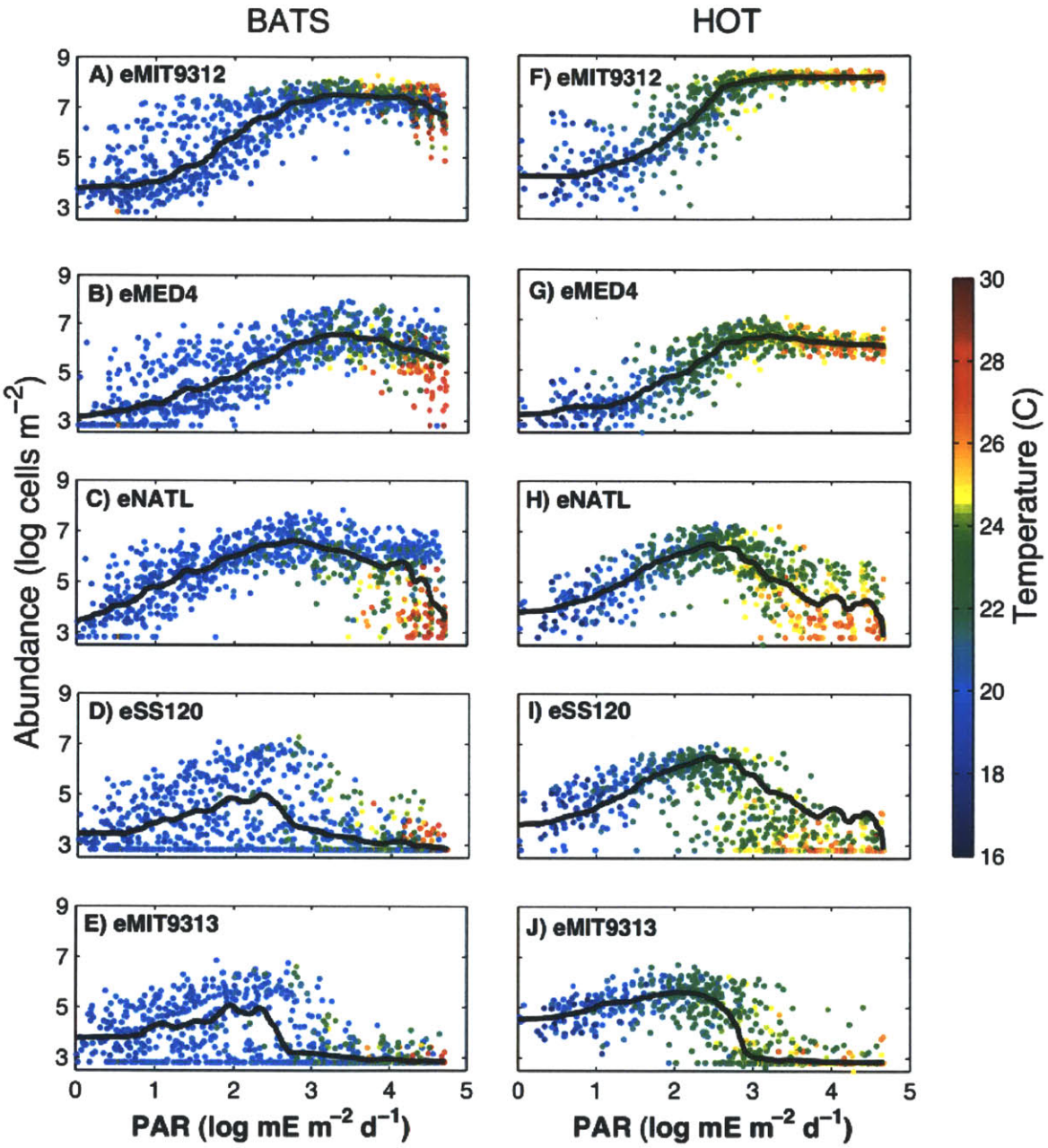


Figure C-2: Ecotype abundance vs. photosynthetically-active radiation at BATS (A-E) and HOT (F-J). Data point colors indicate temperature. Black lines represent a locally weighted regression of the relationship between abundance and irradiance. No profiles were excluded based on mixed layer depth (compare with Fig. 1a,b).

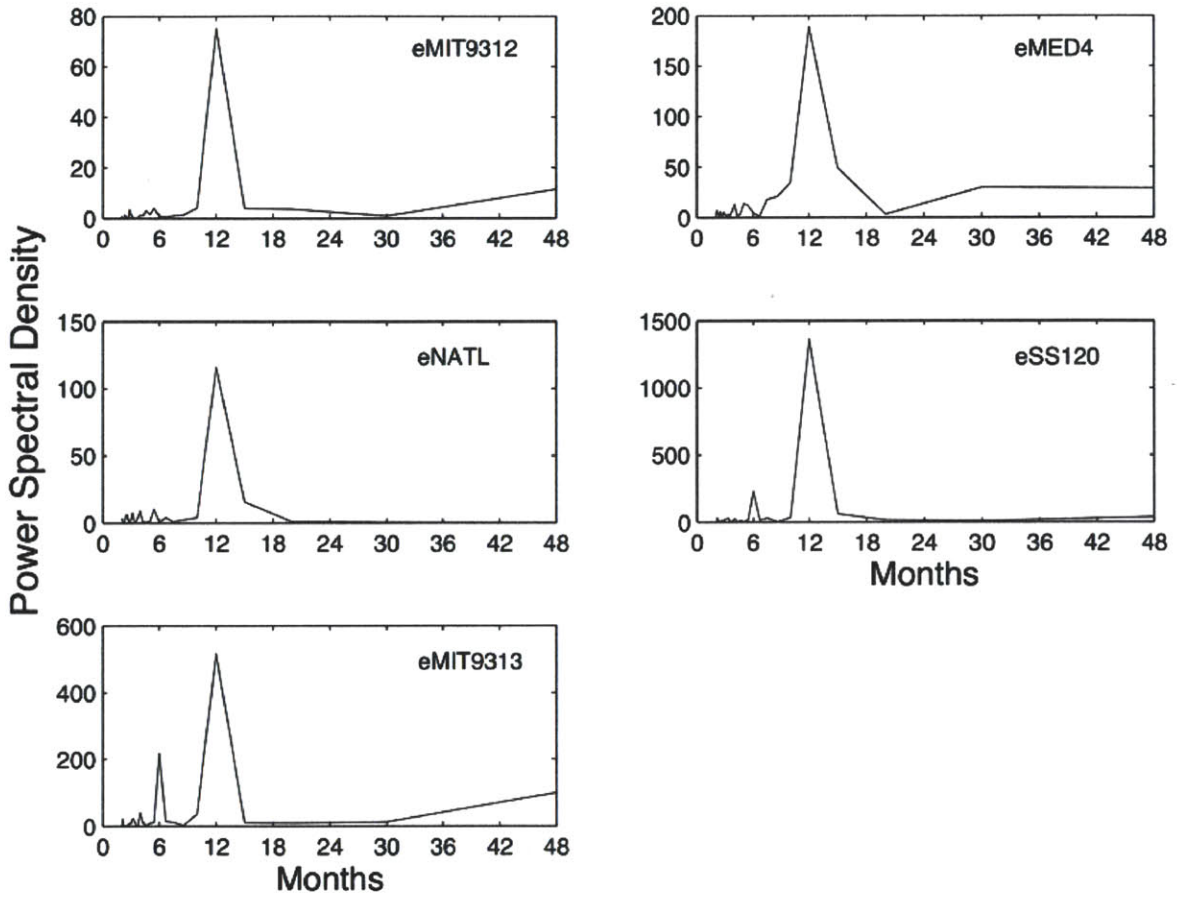


Figure C-3: Spectral analysis of integrated (0-200m) ecotype abundance at BATS. Peaks represent unbiased power spectral density at periods of 1 month.

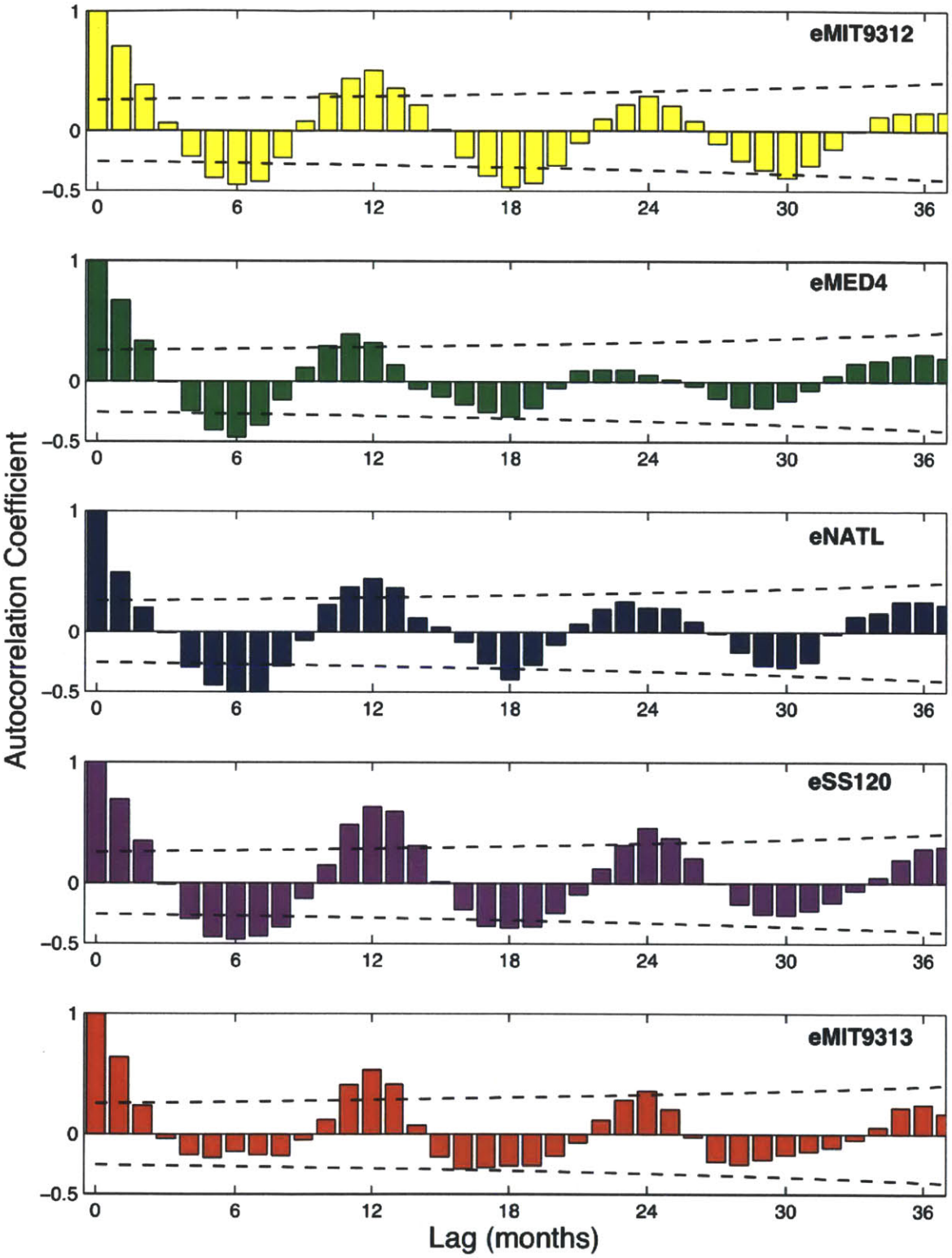


Figure C-4: Autocorrelation of integrated abundance at BATS with lags of one month. Dashed lines represent two standard deviations around a correlation coefficient of 0.

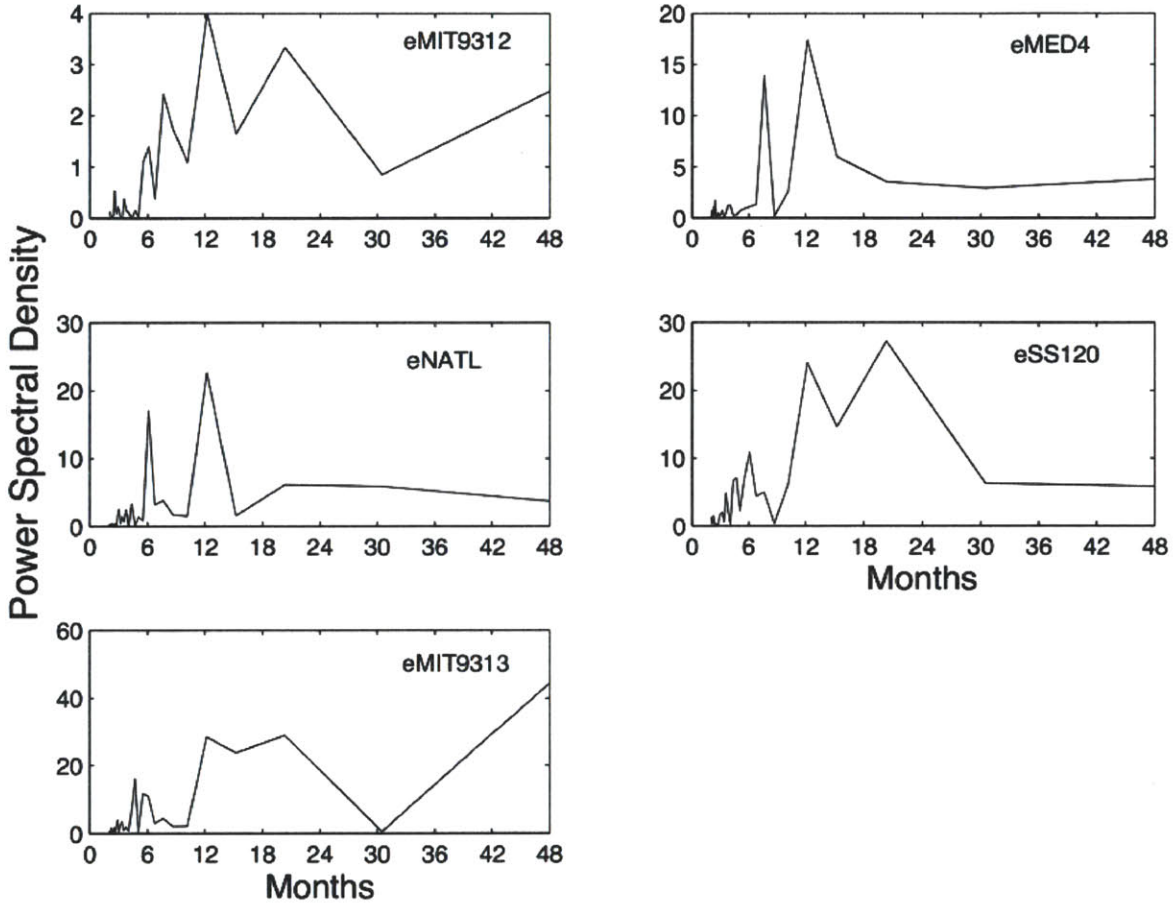


Figure C-5: Spectral analysis of integrated (0-200m) ecotype abundance at HOT. Peaks represent unbiased power spectral density at periods of 1 month.

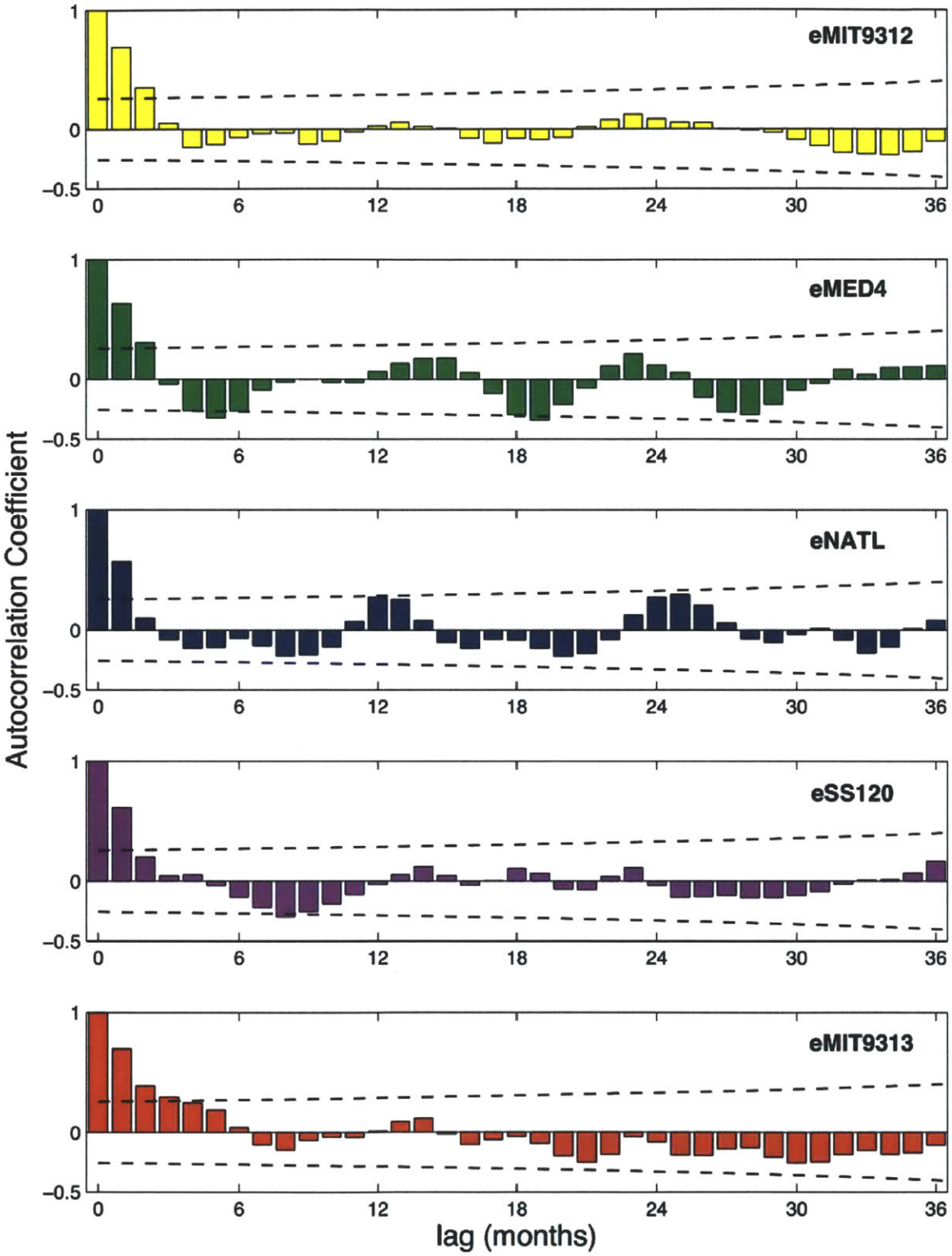


Figure C-6: Autocorrelation of integrated abundance at HOT with lags of one month. Dashed lines represent two standard deviations around a correlation coefficient of 0.

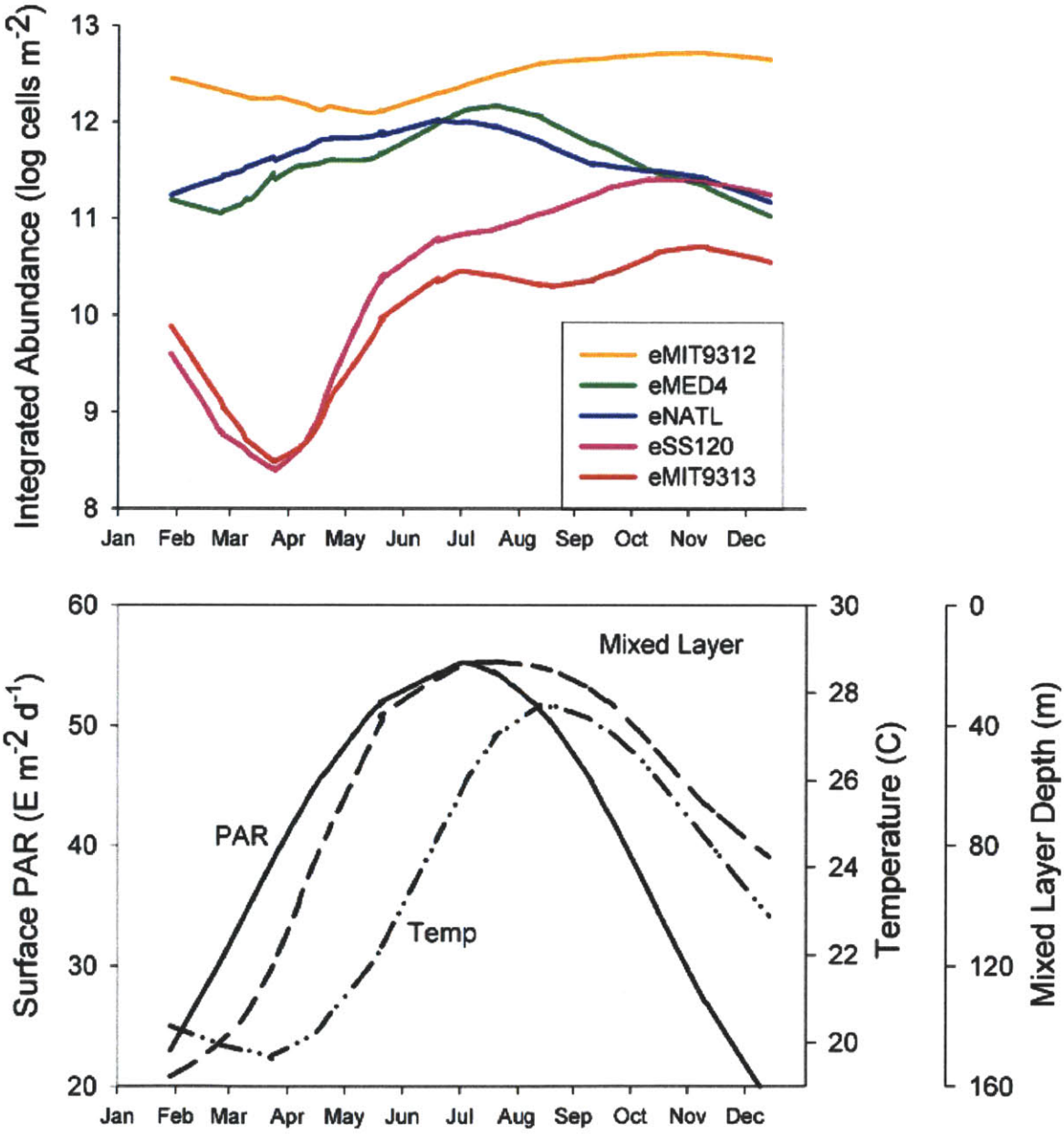


Figure C-7: Annual pattern of integrated abundance, surface temperature, surface light levels, and mixed layer depth at BATS. Data represent a smoothed compilation of all five years.

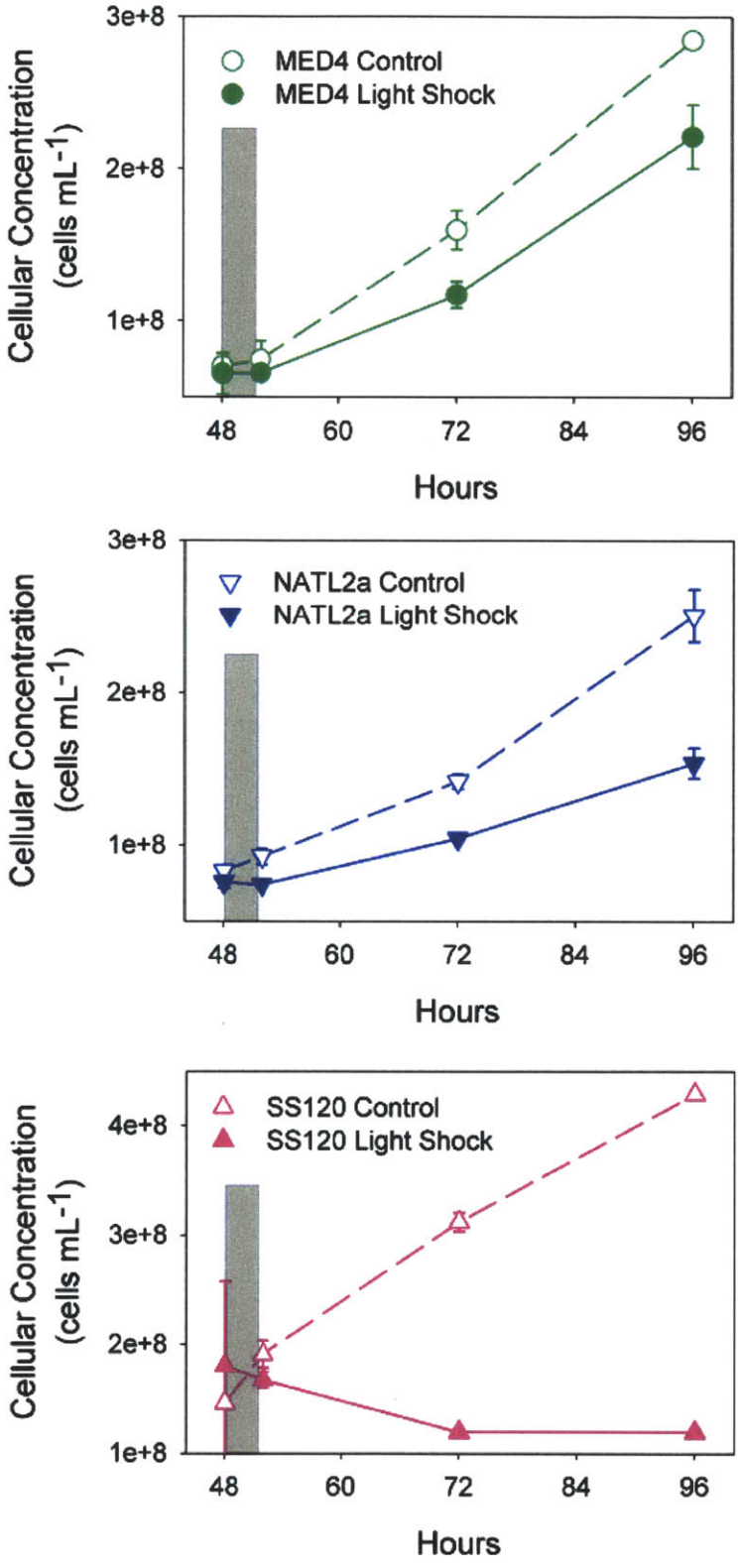


Figure C-8: Response of *Prochlorococcus* strains MED4, NATL2a, and SS120 to light-shock. Cell counts determined by flow cytometry of duplicate cultures acclimated to 35 μmol quanta$\cdot$ $m^{-2}s^{-1}$, exposed to 400 μmol quanta$\cdot$ $m^{-2}s^{-1}$ for four hours (gray area), and returned to 35 μmol quanta$\cdot$ $m^{-2}s^{-1}$. Controls did not experience light shock. Error bars represent one standard deviation.

Appendix D

Preserving and extracting *Prochlorococcus* RNA

D.1 Introduction

The techniques of *Prochlorococcus* transcriptomics are well developed (Lindell et al., 2007; Martiny et al., 2006; Steglich et al., 2006; Tolonen et al., 2006; Thompson et al., 2011; Bagby, 2009; Zinser et al., 2009; Steglich et al., 2010).

However, some concerns remained in the context of a light shock experiment. Concentrating *Prochlorococcus* cells, usually by centrifuge, would take up to 20 minutes from sampling to freezing. This has not been a problem in starvation experiments (Martiny et al., 2006; Tolonen et al., 2006; Thompson et al., 2011; Bagby, 2009) because the experimental cultures remain in their low-nutrient media during centrifugation, while both experiment and control are exposed to the same stress of centrifugation.

However, in a light experiment, centrifugation would mean placing all cultures in the dark for approximately 20 minutes. Given the short half-life of *Prochlorococcus* RNA, this could significantly change the results of the experiment (Steglich et al., 2010). An alternative, which has also been used previously, is filtration. However, setting up a filtration rig to expose cultures to the planned 500 μmol quanta$\cdot$ $m^{-2}s^{-1}$ would be difficult, and previous methods to extract *Prochlorococcus* RNA from a filter involved hot phenol, which we wished to avoid (Steglich et al., 2006).

What was needed was a way to freeze the state of the cell at the time of sampling, for later extraction at leisure. Ideally, this method would work equally well for all *Prochlorococcus* ecotypes,

including eMIT9313 which has previously needed long lysozyme incubations (Tolonen et al., 2006; Martiny et al., 2006). The Chisholm lab has previously used RNAlater (Ambion) to preserve samples from the field, and it naturally stood out as a method to develop further (Frias-Lopez et al., 2008).

D.2 Extraction methods

D.2.1 General concerns

I quickly found that pelleting RNAlater-preserved cultures in the centrifuge was extremely difficult, and had better luck with filtration. The cells remain in RNAlater during filtration, so no change in their RNA content should take place. Surprisingly, RNA yields were better with polyethersulfone (Pall Supor) than with polycarbonate filters.

I tested the stability of RNAlater-preserved cultures and found that there was no noticeable degradation after one week at 4° C. Even after one week at room temperature, RNA was degraded but usable for RT-PCR.

One question was how long to perform a lysozyme incubation for. Previous methods called for a 60-minute incubation, which was effective but could result in degradation of RNA by endogenous RNases which could remain active during the incubation. I saw a reasonable RNA yield after only a 5-minute digest (D-2), and limited all digests to that duration.

Besides the Ambion Mirvana kit, I considered the Zymo Research Quick-RNA kit. This kit is the successor to one that has been used for *Prochlorococcus* microarray experiments (Bagby, 2009), and it is fast and simple while providing good yields. I found it effective and performed some early RT-PCR experiments with it. However, unlike the Mirvana kit it does not retain small RNAs, which would limit it to analyses of gene expression and not of small-RNA regulatory elements.

I also tried the Zymo Research RNA MiniPrep kit. Its simplicity, lack of phenol:chloroform, and ability to capture small RNAs would make it ideal for our purposes. However, in my experience the columns did not adequately separate RNA and DNA. Even after DNase treatment, no-RT controls showed a significant quantity in QPCR. Therefore, I used the Ambion Mirvana kit (Section D.2.3).

D.2.2 RNA extraction from RNAlater-preserved MIT9313 samples with the ZR RNA MiniPrep kit

1. preserve 15 mL culture in 2-3 volumes of cold RNAlater, store at 4° C up to 1 month (ideally less than 1 week)
2. filter on Supor .2 or .4 μM, 25mm filter, proceed or freeze at -80° C
3. prepare 150 μL Tris-HCl (10mM pH 8) + 1 μL Ready-Lyse + 2 μL Superase-In per sample
4. place filter in a 2mL bead beater tube, add 150μL of lysozyme solution to the filter, vortex briefly
5. incubate 5 minutes at room temperature
6. add 600 μL ZR lysis buffer
7. shake 30 seconds at max speed on a bead beater, then cool briefly on ice
8. centrifuge 14000 g 1 minute. Sample should be stable if left on ice here
9. transfer 650 μL to IIIC column (some volume will be lost as it sticks to the filter. May want to draw only 600μL or less to ensure accurate volume)
10. spin 8000 g 30 seconds. Keep flow-through; it's your RNA!
11. add 520 μL (or 0.8 * step 9 volume) and mix
12. transfer half to IIC column. Spin briefly, discard flow-through, and add the rest of your sample to column.
13. spin 14000 g 1 minute, discard flow-through
14. add 400 μL Prep Buffer
15. spin 14000 g 1 minute, discard flow-through
16. add 800 μL Wash Buffer (check that EtOH was added)
17. spin 14000 g 30 seconds, discard flow-through
18. add 400 μL Wash Buffer
19. spin 14000 g 30 seconds, discard flow-through
20. spin again 14000 g 2 minutes to ensure all EtOH is gone
21. add 50 μL nuclease-free water
22. wait 1 minute at room temperature
23. spin 10,000 g 30 seconds
24. check sample on NanoDrop and/or Bioanalyzer and store at -80 C

D.2.3 RNA extraction from RNAlater-preserved MIT9313 samples with the Ambion Mirvana kit

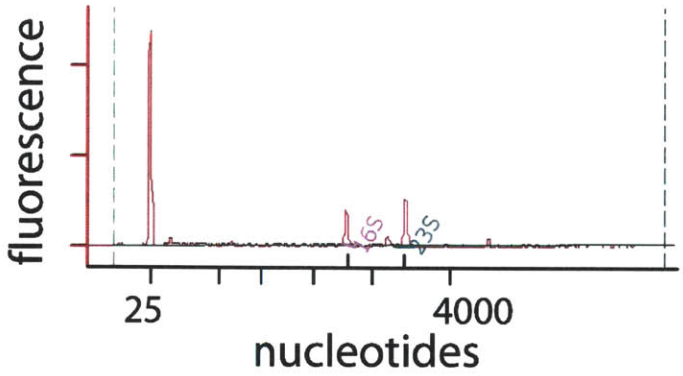
1. preserve culture in 2-3 volumes of cold RNAlater, store at 4 up to 1 month (ideally less than 1 week)
2. filter on Supor .2 or .4 micron, 25mm filter, proceed or freeze at -80° C
3. prepare 150 μL Tris-HCl (10mM pH 8) + 1 μL Ready-Lyse + 2 μL Superase-In per sample
4. place filter in a 2mL bead beater tube, add 150μL of lysozyme solution to the filter, vortex briefly
5. incubate 5 minutes at room temperature
6. add 700 μL Lysis/Binding Buffer
7. shake 30 seconds at max speed on a bead beater, then cool briefly on ice
8. centrifuge briefly
9. transfer 700 μL to new tube
10. add 70 μL Homogenizer Additive
11. vortex 15 seconds
12. on ice 10 minutes
13. add 700 μL acid phenol:chloroform
14. vortex 30 seconds
15. centrifuge 20,000 g for 5 minutes
16. remove aqueous layer (aim for 600 μL, without disturbing interphase or phenol layer) to new tube
17. add 750 μL (1.25 volumes) 100% ethanol and mix
18. transfer half to column, centrifuge 10,000 g 20 seconds, discard flow-through
19. transfer the rest, spin again
20. add 700 μL Wash 1 Buffer, centrifuge 10,000 g 10 seconds, discard flow-through
21. add 500 μL Wash 2/3 Buffer, centrifuge 10,000 g 10 seconds, discard flow-through
22. repeat with Wash 2/3 Buffer
23. centrifuge again 10,000 g 1 minute to remove ethanol traces
24. transfer column to collection tube, add 100 μL elution buffer at room temperature
25. wait 1 minute
26. centrifuge 20,000 g 30 seconds

27. (optional) repeat elution with same or fresh elution buffer
28. check sample on NanoDrop and/or Bioanalyzer and store at -80 C

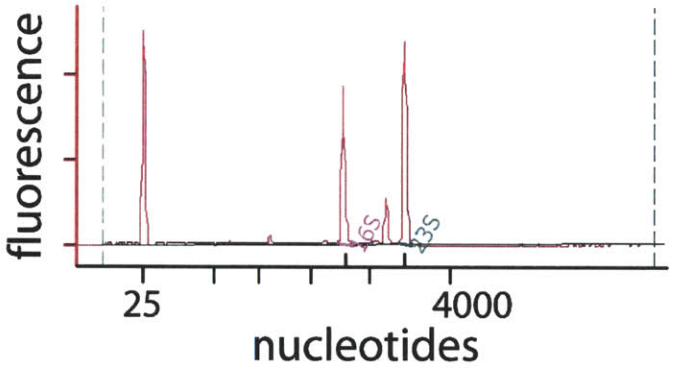
D.3 Results

RNA can be extracted from RNAlater-preserved NATL2Aax with the Zymo Research kits, but only if the samples are first diluted in Tris-HCl, which may simply dilute the RNAlater (Fig. D-1). Lysozyme digestion (which necessarily dilutes the RNAlater in Tris) also helps and may improve yield slightly (Fig. D-1B), but Tris alone is sufficient (Fig. D-1C).

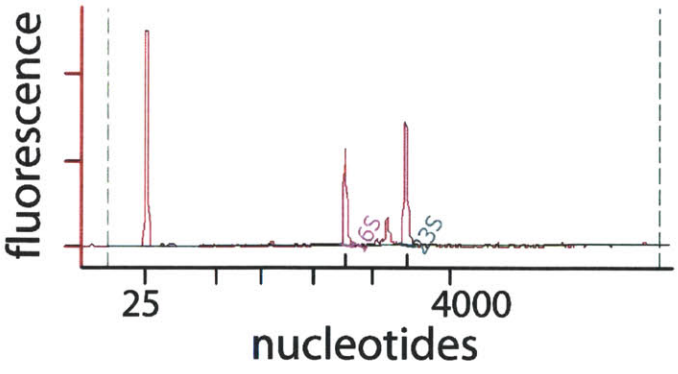
In the case of MIT9313ax, the lysozyme digest was essential prior to using the kit (Fig. D-2), which confirmed past experience with the Ambion Mirvana kit (Tolonen et al., 2006). However, the lysozyme incubation can be very short and can be carried out at room temperature (Fig. D-3), which is different from the established protocol. This finding is useful as it leaves less time for endogenous RNases to re-activate and alter the transcriptome before the cell is lysed.



(a) GTC incubation

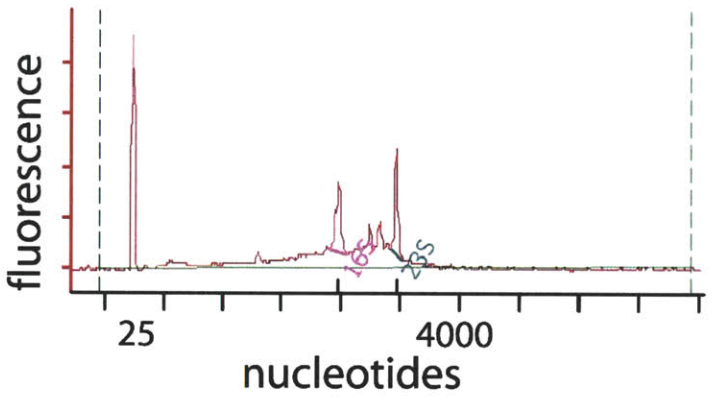


(b) lysozyme incubation

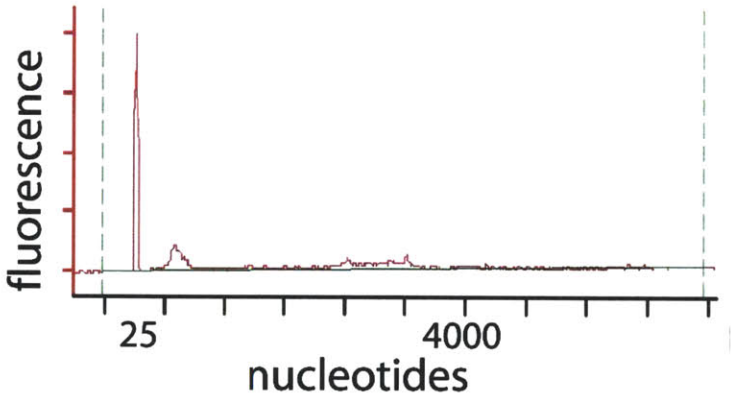


(c) Tris incubation

Figure D-1: Agilent Bioanalyzer traces demonstrating the effect of extraction of RNA from NATL2Aax in RNAlater requires dilution in Tris-HCl. All samples were preserved with RNAlater prior to extraction with the Zymo RNA MiniPrep kit. (A) Used the kit as described, except the sample was incubated in GTC (the first of the kit's reagents) for 10 minutes. (B) Prior to using the kit, the sample was incubated in lysozyme + Tris-HCl. (C) Prior to using the kit, the sample was incubated in 150 μL Tris-HCl only, without lysozyme.

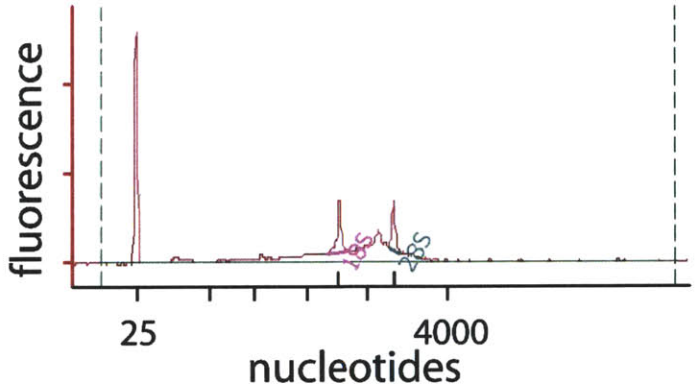


(a) lysozyme incubation

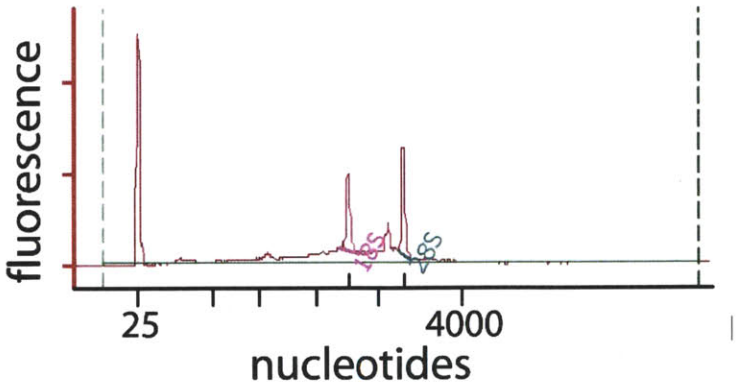


(b) Tris incubation

Figure D-2: Effect of lysozyme on RNA extraction from MIT9313ax using the Zymo Research RNA MiniPrep kit. (A) incubated 30 minutes, 37 ° C in lysozyme + Tris. (B) incubated 30 minutes, 37 ° C in Tris only.



(a) MIT9313ax, 5 minutes lysozyme room temp



(b) MIT9313ax, 20 minutes lysozyme room temp

Figure D-3: Effect of duration of lysozyme digest on MIT9313ax RNA yield. (A) incubated 5 minutes. (B) incubated 20 minutes. Both were incubated at room temperature.

Appendix E

Changes in gene expression during a light shock

E.1 Introduction

Chapter 3 examined the differing physiological responses by *Prochlorococcus* isolates to high-light shocks. In addition, Steglich et al. (2006) examined the global gene expression response to light exposure in *Prochlorococcus* MED4 cells when shifted to moderate light (55 μmol quanta$\cdot$ $m^{-2}s^{-1}$) after 5 hours in darkness. However, we did not know if cells respond the same way at the transcriptional level to shifts from one non-zero light intensity to another, nor how the response of other *Prochlorococcus* isolates might differ. To evaluate the scale and timing of the cellular response to high light, we planned to measure the global transcription response in NATL2Aax and MIT9313ax as well. To help plan this experiment, we targeted a small number of genes with RT-PCR to verify that cells were responding on a transcriptional level, and to determine the timing of that response. Those results are reported here, but the same samples should be used for a whole-genome analysis in the near future.

E.2 Methods

E.2.1 Sampling

Samples were taken from cultures of MED4ax, NATL2Aax, and MIT9313ax that were acclimated to 35 μmol quanta$\cdot$ $m^{-2}s^{-1}$ and shifted to 500 μmol quanta$\cdot$ $m^{-2}s^{-1}$ for 2 hours. For RNA extraction,

15 mL of culture was removed and immediately mixed into 2.5 volumes RNAlater (Ambion) at 4° C. All sampling was performed at 0, 30, 60, 120, and 180 minutes. Experimental cultures were returned to 35 μmol quanta· $m^{-2}s^{-1}$ at 120 minutes; the 180 minute timepoint is after 1 hour of recovery.

RNAlater-preserved samples were filtered the next day on 0.2 micron Supor (Pall) filters and frozen at -80° C. Frozen filters were thawed at room temperature, then coated with 200 μL Tris-HCl (pH 8) and lysozyme. Lysozyme incubation was allowed to proceed 5 minutes at room temperature; this was enough to maximize RNA yield from MIT9313ax (Appendix D).

The RNA lysis buffer from a Mirvana RNA extraction kit (Ambion) was then added. The filter and mixture were shaken on a bead beater for 30 seconds to assist in lysis and removal of cells from the filter. Extraction proceeded according to the manufacturer's protocol. The RNA sample was treated with Baseline-ZERO DNAse (Epicentre) for 20 minutes at 37° C remove contaminating DNA.

E.2.2 Target genes

The selected genes have previously been shown to be upregulated in light shifts in MED4 (Steglich et al., 2006), along with their orthologs in NATL2Aax and MIT9313ax (Table E.1). The work of Steglich et al. (2006) concerned a shift of dark-adapted cells to moderate light, suggesting these genes were candidates to be upregulated in a shift between light levels, as well.

Among the genes chosen was one *hli* operon from each genome. Here it is important to distinguish between core, freshwater cyanobacteria-like *hli* copies and island-borne, phage-like copies (Lindell et al., 2007, Appendix A, Chapter 4). Phage-like *hli* copies were targeted because past experiments have shown these are the only copies to be upregulated by stresses including light (Steglich et al., 2006), phage infection (Lindell et al., 2007, Appendix A), iron starvation (Thompson et al., 2011), and nitrogen starvation (Tolonen et al., 2006). The other two genes selected, PMM1001 and PMM1168, are of unknown function but were chosen because they were among the most upregulated in the MED4 light shift experiment (Steglich et al., 2006), and because they have orthologs in the LL genomes.

E.2.3 RT-PCR

RNA samples were reverse transcribed using the iScript cDNA Synthesis Kit (Bio-Rad), following the manufacturer's protocol. A set of controls were incubated without reverse transcriptase at the

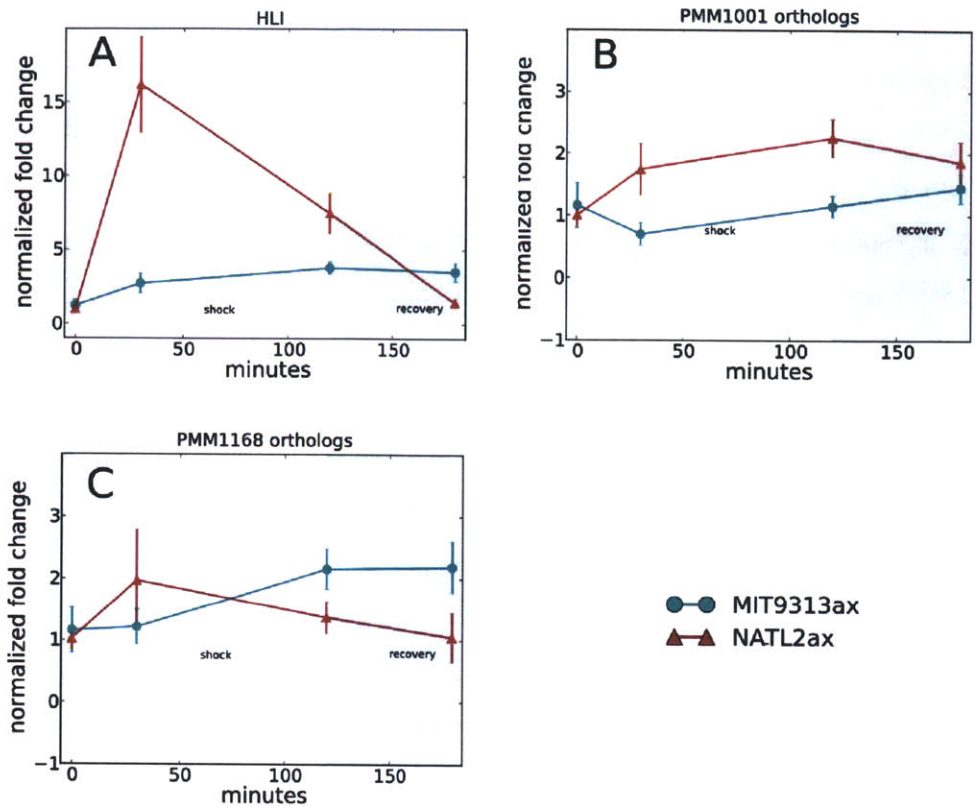


Figure E-1: Changes in expression of targeted genes during a light shock. The shaded duration indicates 35 μmol quanta$\cdot$ $m^{-2}s^{-1}$; the unshaded represents 500 μmol quanta$\cdot$ $m^{-2}s^{-1}$. Normalized fold change is experiment, normalized to *rnpB*, over control, normalized to *rnpB*. (A) One phage-like, multi-copy *hli* gene operon from each isolate. (B) Orthologs of the MED4 PMM1001 gene. (C) Orthologs of the MED4 PMM1168 gene.

same time. Three technical replicates for each sample or genomic standard were subjected to QPCR with the Quantitect SYBR Green PCR Kit (Qiagen) on a CFX96 Real-Time PCR system (Bio-Rad). Expression was normalized to ribonuclease P (*rnpB*), a housekeeping gene whose expression is stable through changes in light level (Mary and Vaulot, 2003; Mary et al., 2004). Expression is reported as fold change experiment/control.

Expression of the *hli* genes exhibits the greatest difference between NATL2Aax and MIT9313ax (Figure E-1). NATL2Aax *hli* expression increases 15-fold in the first 30 minutes, then gradually declines even before being removed from high light. MIT9313ax's response is dampened, but it actually maintains an elevated level 60 minutes into recovery, where NATL2Aax *hli* returns to its pre-shock expression level. Here, the evidence suggests that *hli* expression is part of the primary response to an increased light level in NATL2Aax, and by some models, HLIPs could be responsible for the early quenching of fluorescence (Figure 3-7) (Havaux et al., 2003). In this case, MIT9313ax

may have no primary response, leading to a high level of oxidative stress. That stress may activate *hli* expression and keep it high even into recovery. If so, and if the elevated level at 3 hours indicates MIT9313ax is still experiencing that stress, this could be a deciding factor in the ultimate survival of one strain but not the other.

This delay is especially intriguing as it is duplicated by the response of at least one other gene, MIT9313ax's ortholog of PMM1168. The delays suggest a scenario in which the "truly LL" MIT9313ax cell, exposed to the same stress and possessing many of the same response genes, lacks the regulatory mechanism necessary to respond to that stress as NATL2Aax does. NATL2Aax, in contrast, responds inconsistently at 30 minutes (one of three bottles showed a significant increase in expression), but is clearly back to its initial level during recovery. The PMM1168 ortholog may therefore have little or no role in NATL2Aax's response to a light shock, whereas MIT9313ax does upregulate it. In the case of their orthologs to PMM1001, the difference is again stark, as NATL2Aax upregulates it immediately (if only slightly) and MIT9313ax never does.

This delayed response across putative light response genes may also mirror a previous study of proteases and chaperones, some of which respond quickly in MED4 but remain the same or decrease in transcript level in MIT9313 (Mary et al., 2004). That result, as the ones here, suggests that as different isolates employ different genes in the early minutes of a stress response, they may have less need for other stress response genes later on.

The massive upregulation of *hli* transcripts in NATL2Aax suggests that they are indeed important, but the much smaller, delayed response of the ortholog in MIT9313ax suggests that more separates these isolates' responses than simply the quantity of genes. This raises the question of whether genes' differential regulation, or their complete absence from some strains, plays the greater role in adaptation to diverse environments.

	Gene Name	Locus	MED4 Name	Forward Primer	Reverse Primer
NATL2a	hli	NATL2 10621-601	-	CATGATGGCATTCGTTCTTCT	AGCCATTGAGTCTTTCTGCAA
	rnpB	NATL2_R0025	RNA_42	GCCTAACGCTTAATGGGGATA	GGCTCTCTACAGGGACAAG
	-	NATL2_10331	PMM1001	AGCGAATGGTTGGCTTATTG	AGCGAATGGTTGGCTTATTG
	-	NATL2_10011	PMM1168	CAGGCTGGCAATATCTTCAA	TTTTGTCCAGAATGCTCGATT
MED4	rnpB	RNA_42	RNA_42	TATGTATGGCAAGGGTGCAA	GCCGGGTTCTGTTCATCTTA
	hli		PMM0818-PMM0817	TGACACCTGACGCAGAAAGA	GCCCAACGACCGTTAACTT
	-	PMED4_11261	PMM1001	GGTTCAAACGGTTGGTTGAT	GGGTAAGCCTTGCGTCAATA
	-	PMED4_13331	PMM1168	TTGGGCATGGATTAAACAAA	GGCTGCCACCATTAAACAAT
MIT9313ax	-	P9313_12731	PMM1168	CCTGAGGATATGGCTGAGGA	ACCCCAAGACACCTGTTGAG
	-	P9313_17401	PMM1001	GATCTTGGTCTGCCTCTTGC	TCCCTCCATCTGCATAAAGC
	rnpB	RNA_47		AAGACGAGCTTGGTTGAGGA	CTCTTACCGCACCTTTGCAC
	hli	PMT1154-53		GCGGTCGGAGCATATCTAAC	TTGGCCAGTTCTGCTTCTTT

Table E.1: Genes selected for RT-PCR in the light shock experiment.

Appendix F

Evaluating the *Prochlorococcus* photosystem with the Satlantic FIRe

F.1 Introduction

One simple approach to measuring the health of the photosystem is to measure its capacity to absorb excitation energy without passing it on to the electron transport chain. This has variously been done with DCMU, pump-and-probe, pulse amplitude modulation (PAM), and fast repetition rate fluorometry (FRRF) methods (Röttgers, 2007). While the details vary, all of these methods use either an inhibitor (DCMU) or very high speed to prevent the first transfer of electrons, a reaction that takes on the order of 100-200 μsec. In each of these methods, the fluorescence of a relaxed cell, F_0, with all photosystems open, will be lower than that of a fully excited cell F_m, which releases much incoming energy through fluorescence. The difference between these two readings is called the variable fluorescence F_v. The magnitude of F_v will vary with the number of photosystems or cells in a given culture, so it is almost always normalized against F_m, giving the normalized maximum photosynthetic yield F_v/F_m.

Fluorescence induction and relaxation (FIRe), in its single-turnover induction phase, is similar to FRRF, employing a single light flash (instead of multiple pulses) on the order of 100-200 μs to excite photosystem II (Gorbunov and Falkowski, 2004). Simultaneously, a series of pulses measures the fluorescence of the culture. As in the case of FRRF, this should provide values for F_0 and F_m.

The FIRe instrument is fast and easy to use, but the interpretation of its data can be complex. One problem is that the first 2-3 timepoints are consistently inaccurate, skewing the induction curve

and leading to poor estimates of F_0. For this reason, the Fireworx program includes the option to drop those timepoints entirely (Barnett, 2007). Here I report on the effect of that omission. I also investigate an alternative solution to the same problem: normalizing the first three data points with a chlorophyll standard.

Another parameter reported by Fireworx is σ, a measure of the effective photosynthetic cross section and thus the speed with which the culture reaches F_m. I report very preliminary data concerning this parameter, which could be expected to vary as cells acclimate to different light intensities. The FIRe also reports single-turnover relaxation (STR) and runs a separate multiple-turnover flash (MTF). Those functions were not used in this thesis but may be worth future investigation.

F.2 Methods

F.2.1 FIRe

100 or 200 μL of culture was transferred to a FIRe cuvette and diluted with 1 mL Pro99. Culture was allowed to rest in the dark 5 minutes unless otherwise noted. Results were processed with Fireworx. Fireworx was calibrated against spinach chlorophyll a (Sigma), varying concentrations. Fireworx was modified to save the fitted STI curves as Postscript files (e.g. F-1).

F.2.2 DCMU

DCMU inhibition was used to measure F_m as described by Thompson (2009): 200 μL of culture was transferred to two wells of a clear-bottomed, 96-well plate. To one well, 3 μL ethanol was added. To the other, 3 μL 10mM DCMU in ethanol was added. The plate was placed in light (35 μmol quanta$\cdot$ $m^{-2}s^{-1}$) for 10 minutes, and chlorophyll fluorescence (440 excitation, 680 emission) was read in a Synergy HT plate reader (BioTek).

F.3 Results

F.3.1 Effect of initial timepoints

I found, consistent with the suggestion by the fireworx documentation, that excluding early time-points from the analysis produced better data, especially from highly stressed cells. This problem is especially acute in the case of severely stressed cultures. While such a culture could be expected to

have an F_v/F_m approaching zero, the first two timepoints appear to always be lower than F_m, creating an illusory F_v. Excluding the first two points was sufficient to produce adequate fits (Figure F-1).

If the first timepoints are consistently under-reported, an alternative approach would be to multiply them by a constant factor to compensate. Cell-free chlorophyll can be used to determine the correct scale factors. Chlorophyll ought to present a horizontal line because no energy can be transferred away by water splitting or electron transport ($F_0 = F_m$). Therefore, factors can be estimated that would flatten the line for chlorophyll data. However, in practice the scale of the first point depends both on the total chlorophyll concentration and on the instrument's gain setting (Fig. F-2), so no one set of factors can be used. The second and third points are more consistent, so a hybrid method could be used: drop the first point and scale the second and third (Fig. F-3).

F.3.2 Comparison with the DCMU method

DCMU (3-(3,4-dichlorophenyl)-1,1-dimethylurea) blocks transfer of the excited electron to quinone A, preventing the transfer of energy away from the excited photosystem II. It has previously been used in *Prochlorococcus* as a slower method to measure F_v/F_m (Thompson, 2009).

To reproduce FIRe results, I tested the same culture with the FIRe and DCMU methods (F-4) during a light shock. The results were consistent in that both methods reported falling F_v/F_m values. However, the DCMU method yields a higher overall value for F_v/F_m. Therefore F_v/F_m values determined by different methods should not be compared against each other. Instead, F_v/F_m should be considered a point of comparison between cultures tested in the same way.

F.3.3 F_0 as an alternative measure of bulk fluorescence

F_0, the fluorescence of a fully-relaxed culture, ought to reflect the total amount of chlorophyll exposed to light. It therefore should offer a way to verify measurements taken with the Turner 10-AU fluorometer in Chapter 3. One important difference is that, for FIRe measurements, cultures were rested for 5 minutes in darkness whereas 10-AU measurements were taken immediately, so the comparison could reveal differences if fluorescence declines during that relaxation (due to photosystems' relaxing to an open state).

A plot of F_0 alone from the same 2-hour timecourse discussed in 3 reveals a picture similar to that of the 10-AU data: MED4ax and NATL2ax fluorescence declines within the first 30 minutes, but MIT9313ax fluorescence does not decline until the second hour, and then only slightly (Fig.

have an F_v/F_m approaching zero, the first two timepoints appear to always be lower than F_m, creating an illusory F_v. Excluding the first two points was sufficient to produce adequate fits (Figure F-1).

If the first timepoints are consistently under-reported, an alternative approach would be to multiply them by a constant factor to compensate. Cell-free chlorophyll can be used to determine the correct scale factors. Chlorophyll ought to present a horizontal line because no energy can be transferred away by water splitting or electron transport ($F_0 = F_m$). Therefore, factors can be estimated that would flatten the line for chlorophyll data. However, in practice the scale of the first point depends both on the total chlorophyll concentration and on the instrument's gain setting (Fig. F-2), so no one set of factors can be used. The second and third points are more consistent, so a hybrid method could be used: drop the first point and scale the second and third (Fig. F-3).

F.3.2 Comparison with the DCMU method

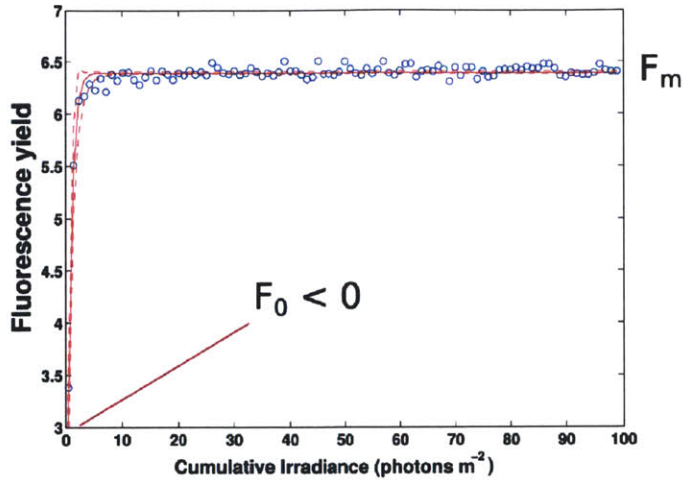
DCMU (3-(3,4-dichlorophenyl)-1,1-dimethylurea) blocks transfer of the excited electron to quinone A, preventing the transfer of energy away from the excited photosystem II. It has previously been used in *Prochlorococcus* as a slower method to measure F_v/F_m (Thompson, 2009).

To reproduce FIRe results, I tested the same culture with the FIRe and DCMU methods (F-4) during a light shock. The results were consistent in that both methods reported falling F_v/F_m values. However, the DCMU method yields a higher overall value for F_v/F_m. Therefore F_v/F_m values determined by different methods should not be compared against each other. Instead, F_v/F_m should be considered a point of comparison between cultures tested in the same way.

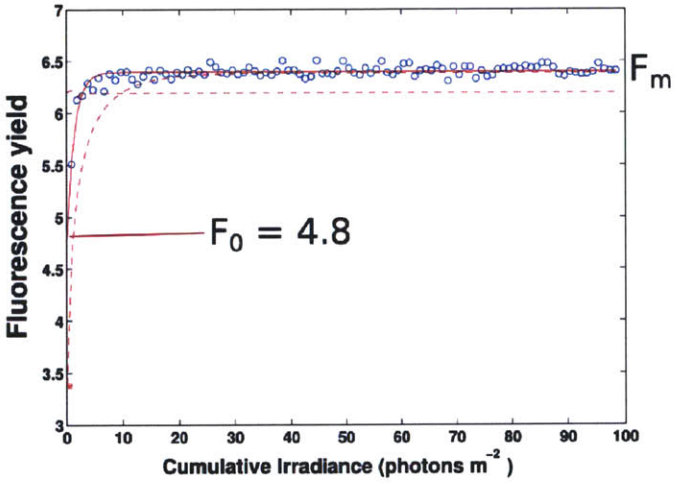
F.3.3 F_0 as an alternative measure of bulk fluorescence

F_0, the fluorescence of a fully-relaxed culture, ought to reflect the total amount of chlorophyll exposed to light. It therefore should offer a way to verify measurements taken with the Turner 10-AU fluorometer in Chapter 3. One important difference is that, for FIRe measurements, cultures were rested for 5 minutes in darkness whereas 10-AU measurements were taken immediately, so the comparison could reveal differences if fluorescence declines during that relaxation (due to photosystems' relaxing to an open state).

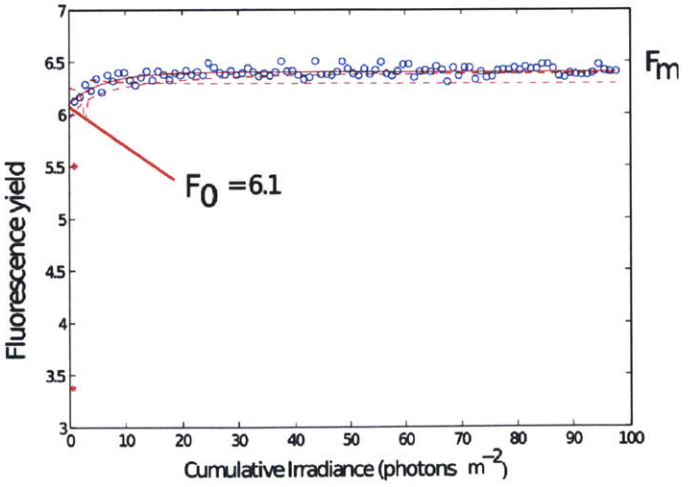
A plot of F_0 alone from the same 2-hour timecourse discussed in 3 reveals a picture similar to that of the 10-AU data: MED4ax and NATL2ax fluorescence declines within the first 30 minutes, but MIT9313ax fluorescence does not decline until the second hour, and then only slightly (Fig.



(a) all data points



(b) excluding first data point



(c) excluding first two data points

Figure F-1: Effect of excluding or including the first 1 or 2 data points on the fit of data from NATL2Aax. Cultures were acclimated to 35 μmol quanta$\cdot$ m^{-2}s^{-1} and shifted to 400 μmol quanta$\cdot$ m^{-2}s^{-1} for four hours before FIRe sampling.

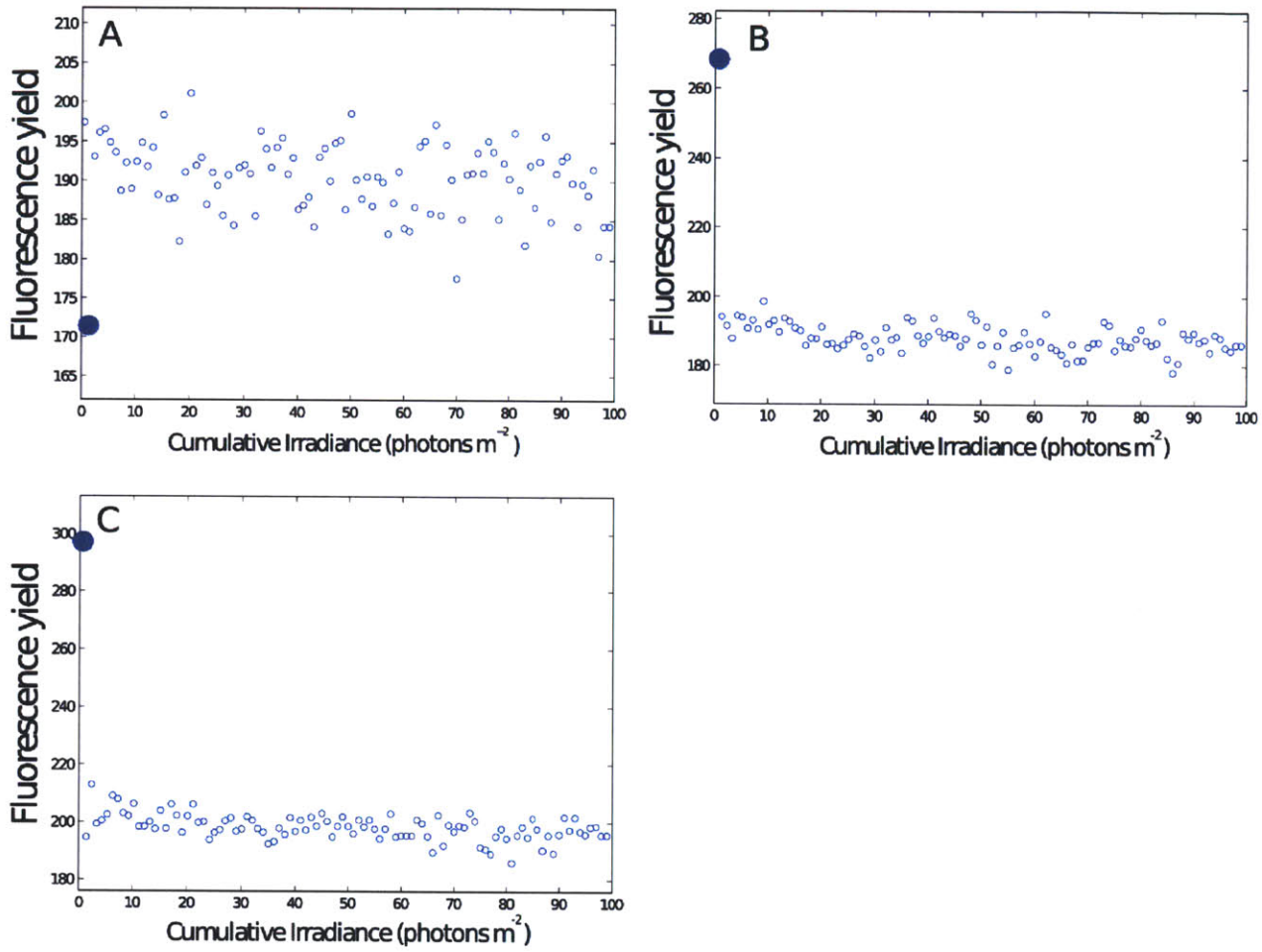


Figure F-2: Normalization of cell-free chlorophyll data by scaling first three datapoints. First three datapoints are scaled by factors of (3.45, 1.15, 1.03). (A) 250 pM spinach chlorophyll a, FIRe gain 1500. First data point is emphasized. (B) 250 pM spinach chlorophyll a, FIRe gain 2000. (C) 500 pM spinach chlorophyll a, FIRe gain 1500.

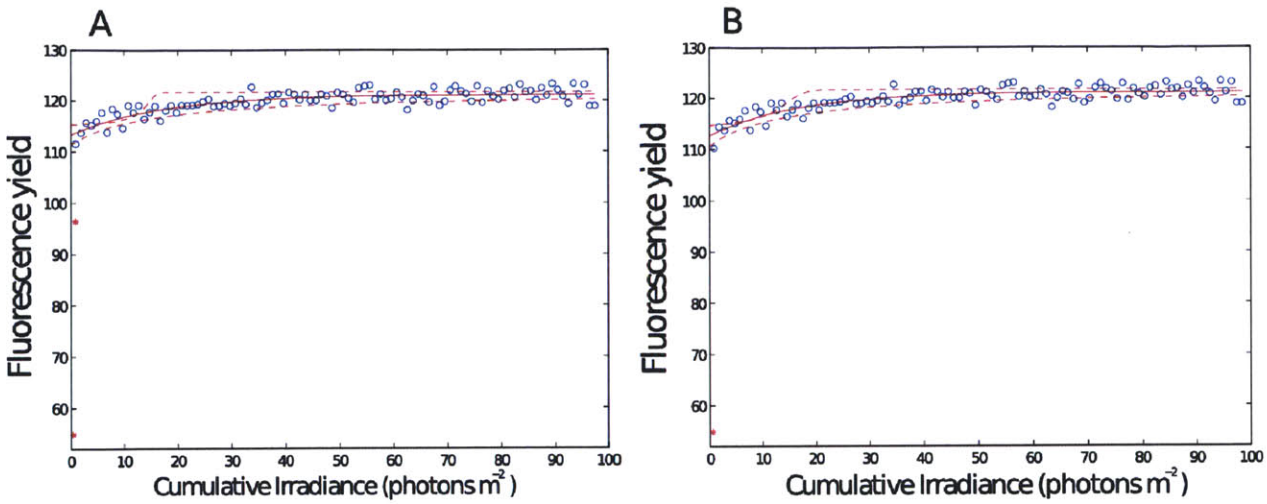


Figure F-3: Normalization of MIT9313ax data using two methods. MIT9313ax was acclimated to 35 μmol quanta$\cdot$ m^{-2}s^{-1} and shifted to 400 μmol quanta$\cdot$ m^{-2}s^{-1} for 4 hours before FIRe data was collected. (A) Dropping first two datapoints, as in Fig. F-1C. (B) Dropping first datapoint and scaling second and third by (1.15, 1.03).

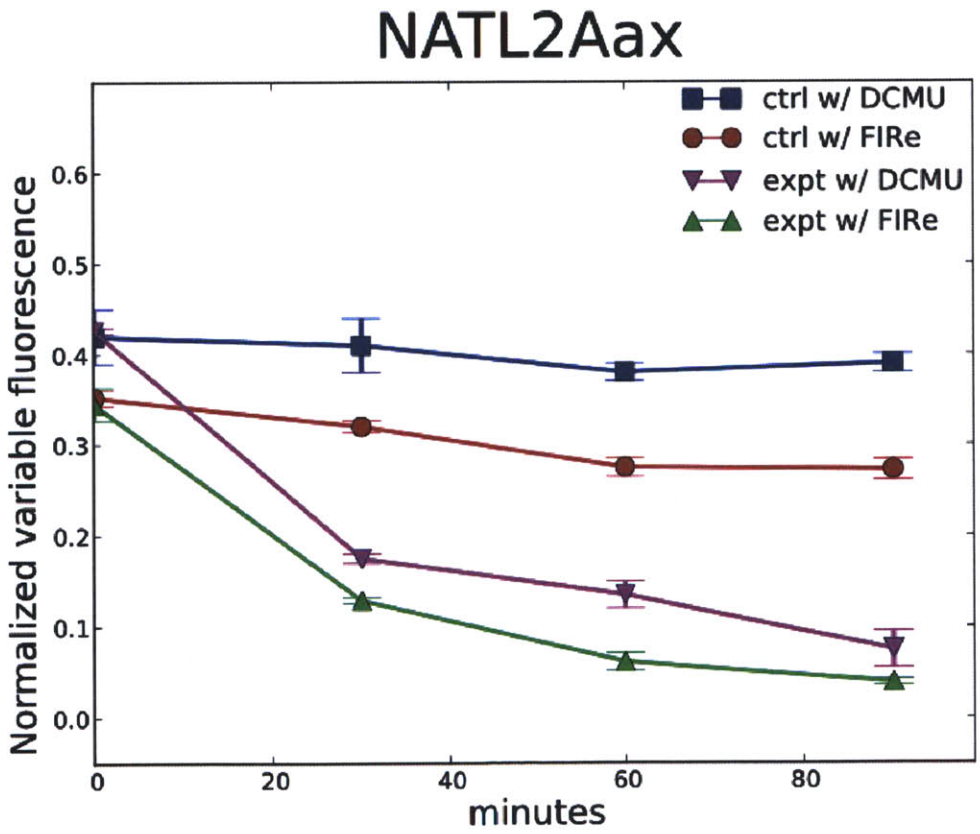


Figure F-4: Comparison of FIRe and DCMU measurements of F_v/F_m. NATL2Aax cultures were acclimated to 35 μmol quanta$\cdot$ m^{-2}s^{-1}, and experimental cultures were shifted to 400 μmol quanta$\cdot$ m^{-2}s^{-1} starting at time 0. Duplicate samples were taken and processed using the FIRe or DCMU methods as described in the methods section.

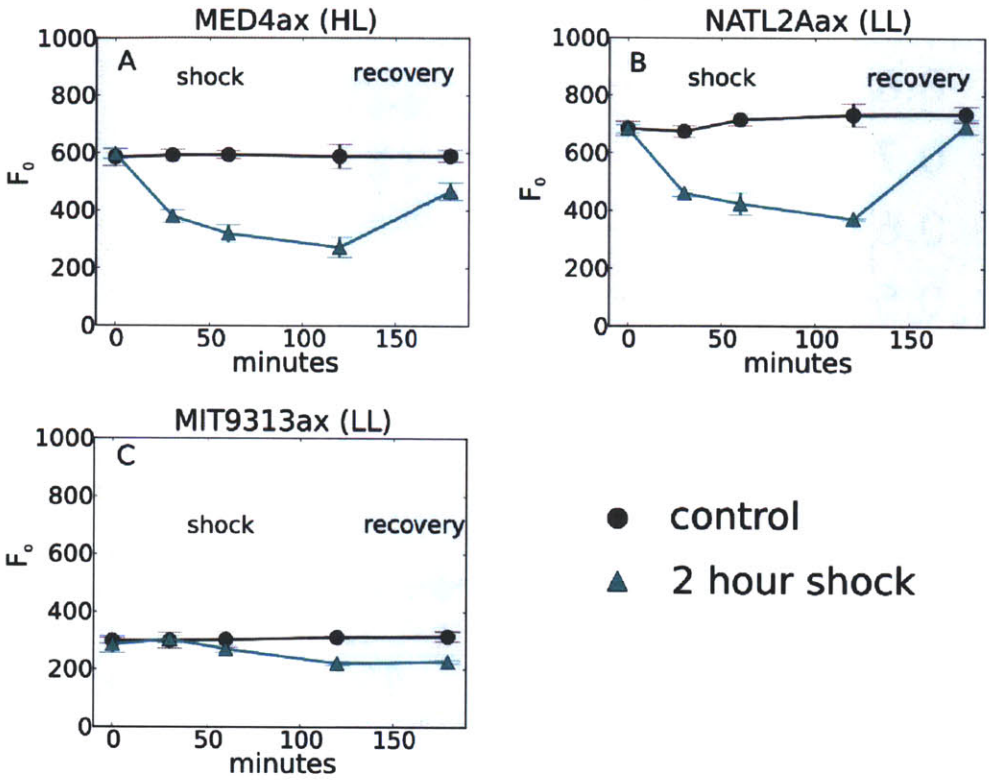


Figure F-5: F_0 (arbitrary units) of *Prochlorococcus* cultures during a 2-hour light shock. Cultures were acclimated to 35 μmol quanta· $m^{-2}s^{-1}$ and shifted to 500 μmol quanta· $m^{-2}s^{-1}$ for 2 hours. They were then returned to 35 μmol quanta· $m^{-2}s^{-1}$. At each timepoint, samples of 200 μL were taken, diluted in 1 mL Pro99, and rested in darkness 5min. They were tested on the Satlantic FIRe with a single turnover flash (STF) duration of 150 μs. (A) MED4ax (B) NATL2Aax (C) MIT9313ax.

F-5). MED4ax and NATL2ax also recover their fluorescence much more quickly than MIT9313ax does in the relaxation phase, again suggesting a reversible quenching mechanism is responsible for the decline. One slight, but potentially important, difference is that MIT9313ax fluorescence does not increase, either, in that first hour, which had been seen with the 10-AU (Fig. 3-7C). This could be explained as an effect of that 5 minutes' relaxation that the FIRe samples were given.

F.4 Future directions

F.4.1 Comparisons within the HL clade

This work has focused on the differences between the clades, particularly HL against LL and eNATL against other LL ecotypes. However, there may be additional differences within those ecotypes. The observation that the HL MIT9301 isolate has only 15 *hli* gene copies (barely more than SS120's

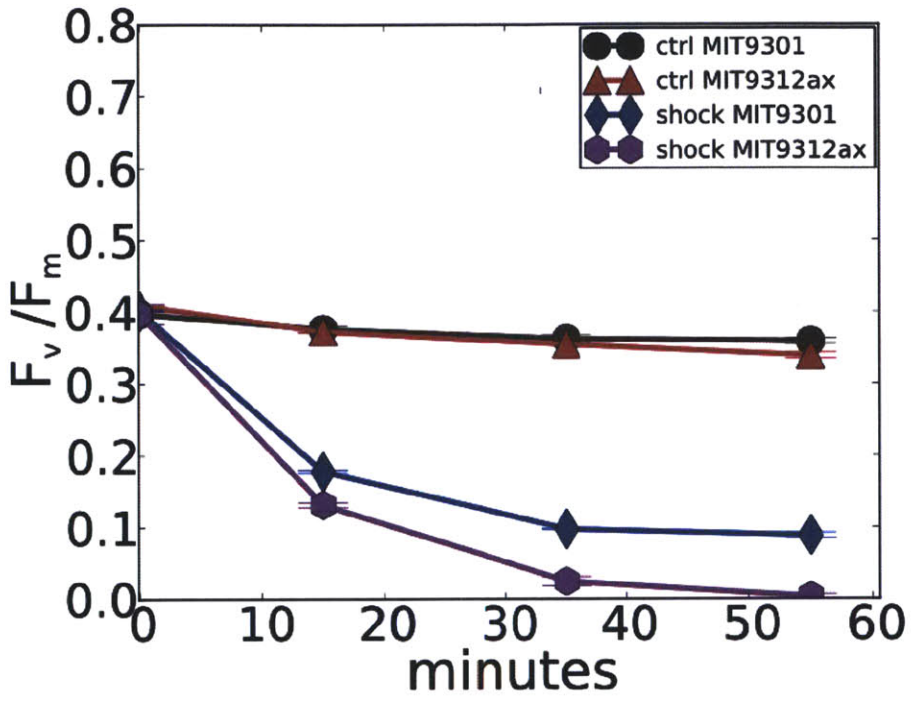


Figure F-6: Response of HL isolates MIT9301 and MIT9312ax to a light shock. Cultures were acclimated to 35 μmol quanta$\cdot$ $m^{-2}s^{-1}$ and shifted to 650 μmol quanta$\cdot$ $m^{-2}s^{-1}$ at 0 minutes. FIRe samples were taken as described above.

12) suggested it might be more susceptible than other HL isolates to a light shock.

In a test of MIT9301 and MIT9312ax (24 *hli* copies), 35 to 600 μmol quanta$\cdot$ $m^{-2}s^{-1}$, the opposite effect was seen: MIT9312ax F_v/F_m was effectively zero after one hour, but MIT9301 F_v/F_m dropped more slowly and leveled off at 0.1 (Fig. F-6). This could also be a result of the lack of heterotrophs in co-culture with MIT9312ax, so comparison with a non-axenic MIT9312 culture would be a necessary first step in following up.

F.4.2 MIT9313 and σ

Further, it appears that MIT9313ax has acclimative abilities not common to the other isolates: σ changes within days of transferring to a new light level. No other LL or HL isolate has been seen to share this ability, which is surprising given the HL and NATL isolates' greater tolerance of high light.

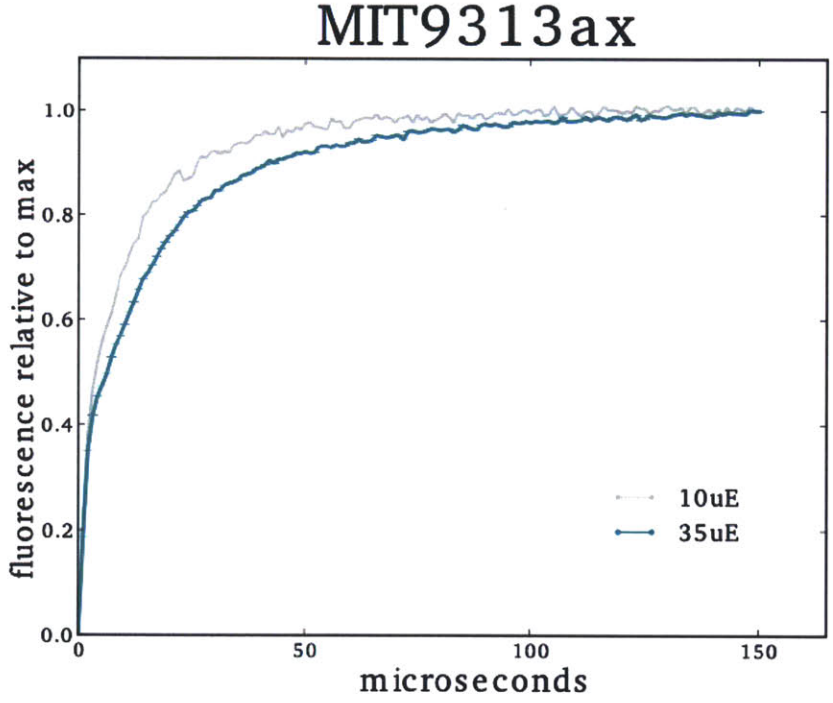


Figure F-7: The increase in fluorescence over the single-turnover induction of MIT9313ax, for cultures adapted to low or moderate light. Separate cultures acclimated to 10 or 35 μmol quanta$\cdot$ $m^{-2}s^{-1}$ were tested in the FIRe. The trace represents the raw data output from the FIRe, before fitting curve parameters. The fluorescence is normalized to the maximum value seen during the timecourse (F_m).

Appendix G

Other *Prochlorococcus* fosmids from HOT Station ALOHA

Gregory C. Kettler, Steven Biller, Jessie Thompson, Maureen Coleman, and Sallie W. Chisholm

In selecting fosmids to sequence in Chapter 4, the primary goal was to capture one eNATL island region with sufficient depth for meaningful analysis. However, the attempt was also made to sample multiple island regions from both HL and LL genomes. Not enough fosmids covered these other regions to answer questions posed in Chapter 4, but they may be of future interest. They are listed in Table G.1, and some alignments against *Prochlorococcus* genomes are presented below.

Most of these fosmids lack *hli* genes. In some cases (Fig. G-5), their comparison against *Prochlorococcus* captures no recombination or gene gains or losses. In others (Fig. G-3), recombination is more extensive than in the region discussed in Chapter 4, suggesting that these regions are more recently changed. That extensive recombination explains much of the difficulty of selecting fosmids that might contain homologous regions, and also points to the urgency of sequencing additional whole genomes to gain a more complete understanding of the recent changes in these highly active genomic regions.

name	depth (m)	date	best match genome	*hli* genes	comment
BYAH6690	125	9-March-2006	AS9601	yes	
BYAF1283	25	9-March-2006	AS9601	yes	incompletely assembled
BYAH19564	125	9-March-2006	MIT9303	no	
BYAH14728	125	9-March-2006	MIT9303	no	
BYAH1163	125	9-March-2006	MIT9303	no	
FNFS10675	110	19-Oct-2006	NATL1A	no	extensive recombination/island
FNFS8854	110	19-Oct-2006	NATL1A	no	extensive recombination/island
FNFS10325	110	19-Oct-2006	NATL1A	no	extensive recombination/island
FNFS15319	110	19-Oct-2006	NATL1A	no	extensive recombination/island
FNFS14460	110	19-Oct-2006	NATL1A	no	extensive recombination/island
FNFS10473	110	19-Oct-2006	NATL1A	yes	extensive recombination/island
FNFS689	110	19-Oct-2006	NATL1A	no	extensive recombination/island
FNFS6174	110	19-Oct-2006	NATL1A	yes	extensive recombination/island
BYAH10984	125	9-March-2006	NATL1A	no	extensive recombination/island
BYAG11462	75	9-March-2006	NATL2A	no	
BYAH15776	125	9-March-2006	NATL2A	no	
BYAG5655	75	9-March-2006	NATL2A	no	
FNFS6678	110	19-Oct-2006	NATL2A	no	
FNFS1285	110	19-Oct-2006	NATL2A	no	
FNFS5150	110	19-Oct-2006	NATL2A	no	
FNFS1236	110	19-Oct-2006	NATL2A	no	
FNFS7944	110	19-Oct-2006	NATL2A	no	
BYAH5143	125	9-March-2006	NATL2A	no	
FNFS835	110	19-Oct-2006	NATL2A	no	
FNFS5079	110	19-Oct-2006	NATL2A	no	
FNFS10147	110	19-Oct-2006	NATL2A	no	
FNFS10708	110	19-Oct-2006	NATL2A	no	
FNFS9086	110	19-Oct-2006	NATL2A	no	
BYAG20101	75	9-March-2006	NATL2A	no	
BYAH17068	125	9-March-2006	NATL2A	no	
BYAH11008	125	9-March-2006	NATL2A	yes	region discussed in Chapter 4
BYAH11841	125	9-March-2006	NATL2A	yes	region discussed in Chapter 4
FNFS10637	110	19-Oct-2006	NATL2A	yes	region discussed in Chapter 4
FNFS415	110	19-Oct-2006	NATL2A	yes	region discussed in Chapter 4
FNFS4828	110	19-Oct-2006	NATL2A	yes	region discussed in Chapter 4
FNFS4961	110	19-Oct-2006	NATL2A	yes	region discussed in Chapter 4
FNFS7206	110	19-Oct-2006	NATL2A	yes	region discussed in Chapter 4
FNFS7728	110	19-Oct-2006	NATL2A	yes	region discussed in Chapter 4
FNFS2462	110	19-Oct-2006	NATL2A	yes	region discussed in Chapter 4

Table G.1: Fosmids sequenced in this study. For each fosmid, the date and depth of its originating sample is given. The genome to which that fosmid is most similar is noted. If a fosmid contains any *hli* genes, that is noted.

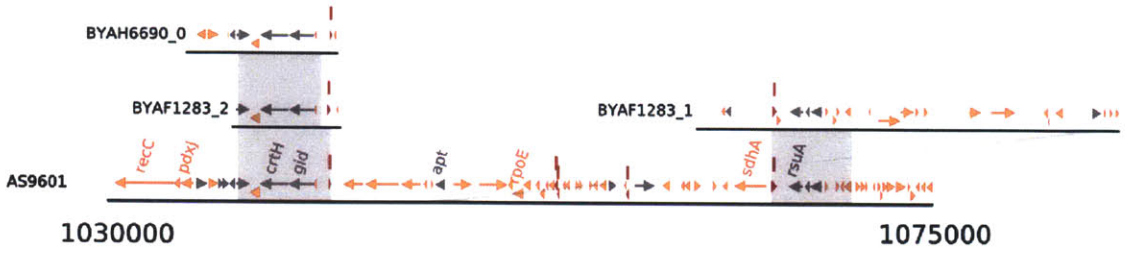


Figure G-1: Fosmids similar to AS9601. The core genome genes are colored black, the flexible genome orange, and *hli* genes, if present, are red. Shaded regions are BLAST alignments.

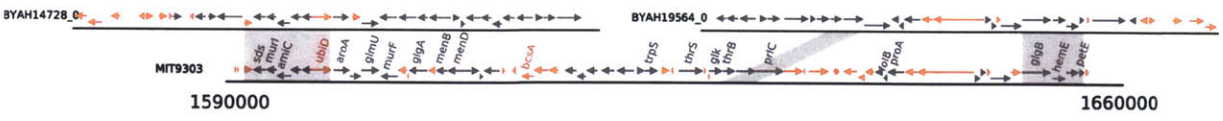


Figure G-2: Fosmids similar to MIT9303.

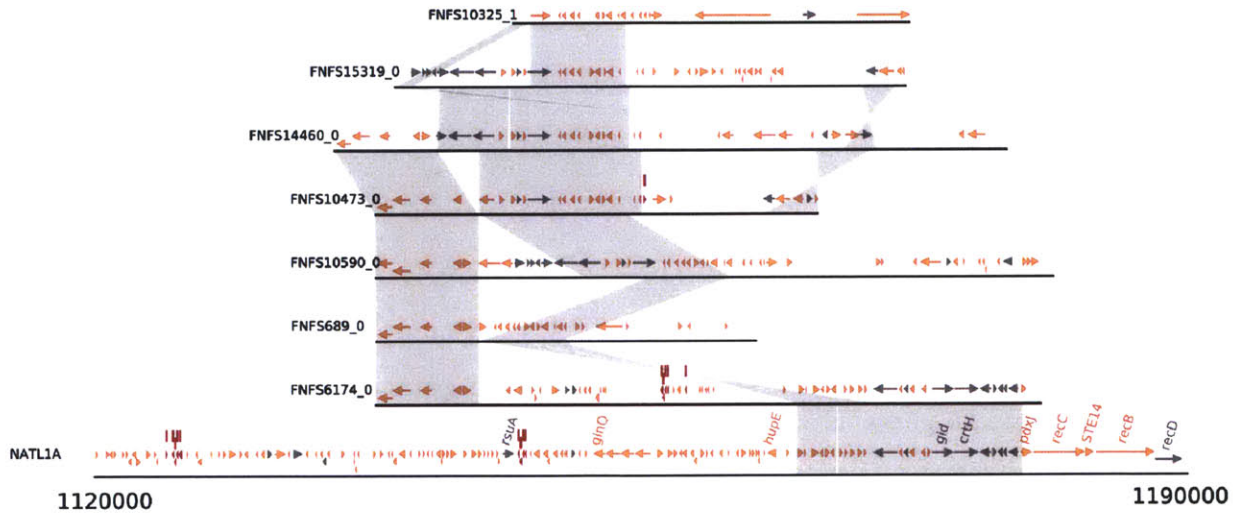


Figure G-3: Fosmids similar to NATL1A.

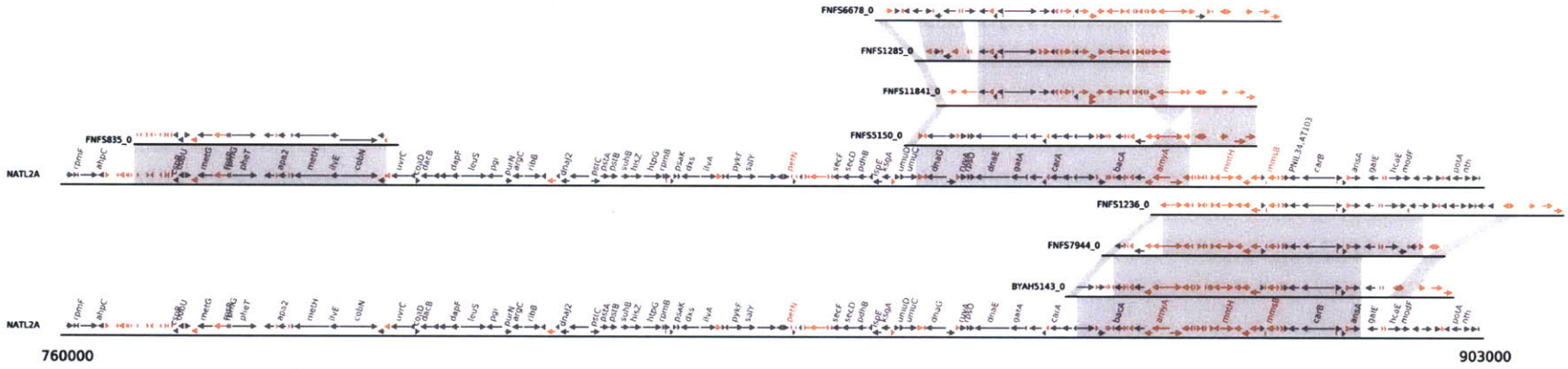


Figure G-4: Fosmids similar to NATL2A.

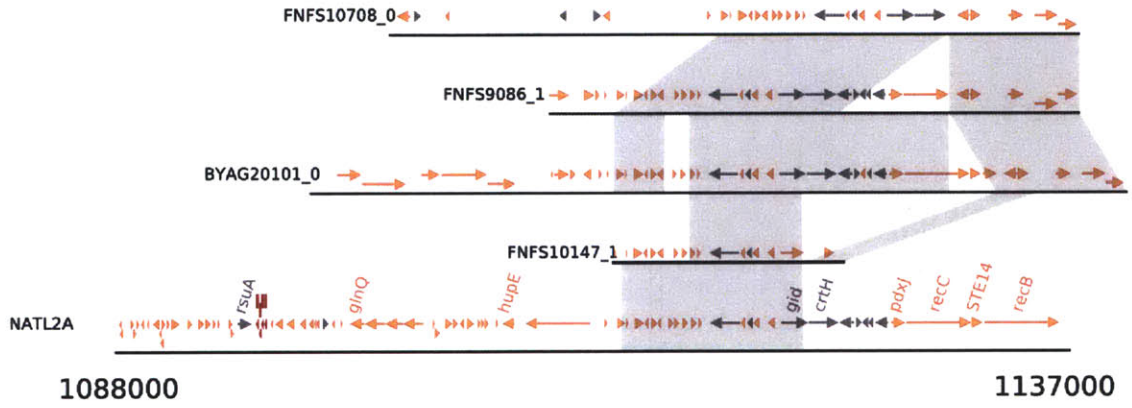


Figure G-5: Fosmids similar to NATL2A.

Appendix H

Using HMMER to identify and classify HLIPs

H.1 Introduction

Because of their similarity to a well-conserved helix in some eukaryotic light-harvesting proteins, high light inducible proteins contain a motif that can be used to identity them (Dolganov et al., 1995; Funk and Vermaas, 1999). In past *Prochlorococcus* studies, it has been sufficient to count matches against the motif, AExxNGRxAMIGF, while accepting those ORFs that match more than 6 conserved residues (Bhaya et al., 2002; Lindell et al., 2004). However, the much greater dataset of the Global Ocean Survey (GOS) requires another method (Rusch et al., 2007). Using the 6-residue cutoff yields a large number of false positives, proteins that on manual inspection do not resemble HLIPs (Fig. H-1A). Increasing the cutoff to 7 excludes many likely HLIPs (Fig. H-1B).

For that reason I used a hidden Markov model (HMM) method with HMMER (Finn et al., 2010). In their application, hidden Markov models are similar to position-specific scoring matrices (PSSMs): a multiple alignment is constructed from a training set, that multiple alignment is used to build a model, and any unknown sequence is tested against that model. However, HMMs also consider the preceding sequence in assigning the probability of a particular residue's occurrence in the next spot (Durbin et al., 1998). PSSMs and HMMs should both be expected to be more sensitive and specific than the motif search described above.

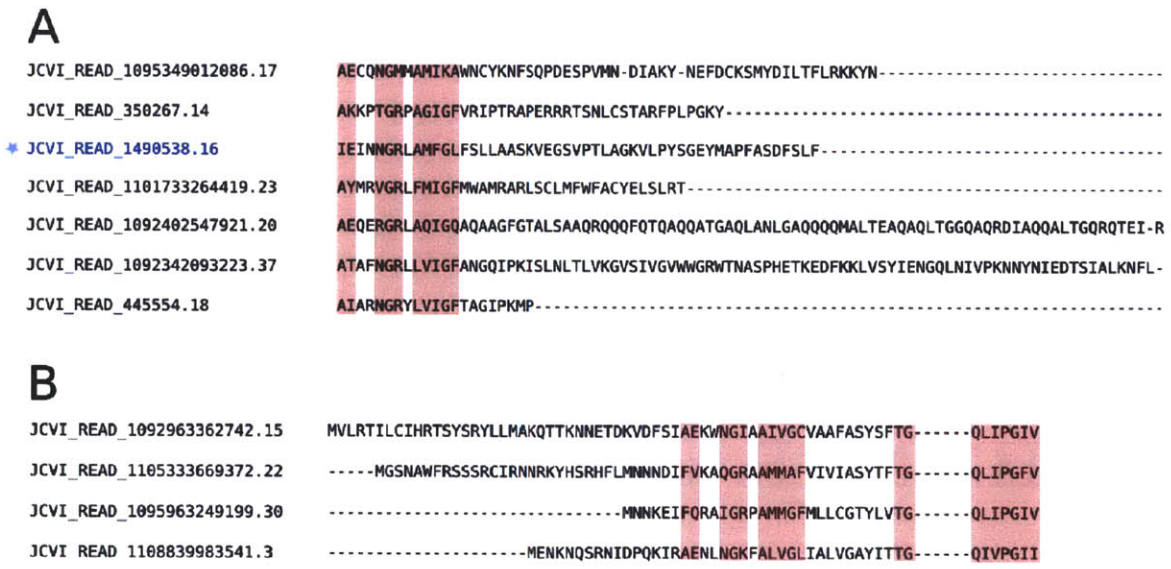


Figure H-1: Mis-identified HLIP sequences using a simple motif search (AExxNGRxAMIGF) in GOS. That motif (and TGQIIPGxF, where applicable) are highlighted. (A) False positive sequences that matches a cutoff of 7 out of 10 residues. The starred sequence matches a eukaryotic light harvesting protein (Fig. H-2). (B) True HLIP sequences that are missed using the cutoff of 7.

H.2 Method

The HMMER package (hmmbuild, hmmsearch) was used (Finn et al., 2010). The outline is as follows:

generate alignments in Stockholm format
for each alignment **do**
  hmmbuild modelname.hmm alignment.sto
end for
for each alignment **do**
  hmmsearch cutoff hmm modelname.hmm queryfile.fasta > modelname.hits
end for
for each query sequence **do**
  determine model with highest domain score and assign to that category
end for

Strictly for ease of alignment, the training set of phage-like, multi-copy HLIPs was broken into five subsets and each was used to train a different model. However, HLIPs matching any of those models were treated equivalently, as one group of phage-like, multi-copy HLIPs. Only one alignment was necessary for the freshwater cyanobacteria-like, core HLIPs.

Figure H-2: Alignment of an ORF, from GOS read JCVI_READ_1490538 (top), against a chloroplast light harvesting protein from *Karlodinium micrum* (GenBank Accession ABV22208, bottom).

H.3 Eukaryotic genes

Because eukaryotic chlorophyll A/B-binding proteins (CAB proteins) and one-helix proteins (OHPs) share the same motif as the HLIPs (hence the alternative name, small CAB-like protein or SCP) some eukaryotic genes may be identified here as well. The approach should be more effective than the simple motif search at excluding these sequences. For example, the ORF JCVI_READ_1490538.16, which is identified by the simple motif search (Fig. H-1B), is arguably a correct identification: it does contain a CAB domain and it is highly similar to a light-harvesting complex protein from the dinoflagellate *Karlodinium micrum* (Fig. H-2). It is not, however, a cyanobacterial HLIP and is not called as such by the HMMER method. It is possible that other eukaryotic proteins or other false positives are still being called by the HMMER approach, in which case the training alignments should be updated to exclude them.

HMMER reports some ORFs to have multiple CAB domains. These were rejected as they are not likely to be true cyanobacterial HLIPs but could be eukaryotic 3-helix CABs.

Appendix I

Genes that may differentiate the HL and LL ecotypes

I.1 Introduction

Many of the preceding chapters are built on the hypothesis that HLIPs, and particularly the number of *hli* copies, differentiate eNATL cells from other LL cells in terms of their ability to recover from high light shocks. It is important to note, however, that the differences between a HL and any LL cell (including an eNATL cell) are more extensive than this one gene family. From Supplementary Table S6 of Chapter 2 (available online at `http://www.plosgenetics.org/article/info:doi/10.1371/journal.pgen.0030231`), one can obtain a list of 99 orthologous groups that are found in all HL isolates but not in any LL isolate. These are candidates for determining the ability of HL cells to grow in sustained high light, which is a selective parameter different from the brief, high light shocks that HLIPs may defend against. Unfortunately, the overwhelming majority of these genes have no hypothetical functional assignments, so their sequences alone will not reveal which are responsible for growth in high light, or how.

One challenge in discussing ecotype-defining genes, however, is in deciding whether to separate them from other, laterally transferred genes that are gained or lost even in comparisons between the most closely related sequenced isolates (that is, between the leaves of the tree). The model of *Prochlorococcus* evolution holds that the divergence from LL to HL, and later between two major HL clades, took place only once, so that their defining traits (light adaptation and temperature adaptation) are monophyletic. But other traits, such as the possession of some nutrient acquisition

genes, are scattered across multiple clades (Martiny et al., 2009b). It is well understood that those nutrient-acquisition genes are located in islands, consistent with the degree of lateral gene transfer that would be required to so scatter them across clades.

It is reasonable, then, to conclude that all non-monophyletic *Prochlorococcus* traits are defined by island genes. But monophyletic traits may also be defined by islands. If eNATL's phage-like *hli* genes, arguably an ecotype-defining gene family, are always located in islands (Chapter 4), other ecotype-defining genes could be located in other genomic islands. Alternatively, because the HL trait arose only once, it could be expected that HL-defining genes integrated among the core genes, and were subject to relatively little subsequent recombination. Given the set of 99 orthologous groups mentioned above, this can be tested.

I.2 Results

The locations of these 99 genes were plotted against the MED4 genomic islands (Fig. I-1). The plot clearly shows that their locations are consistent with MED4's islands. This suggests that the genes that define the HL ecotype are part of the islands of those genomes, even though that ecotype's traits are stable: no descendant of the HL LCA has been seen to exhibit a more LL-like phenotype. This would be similar to the model of eNATL put forth in Chapter 4: a large clade whose defining phenotype is attributed to a conserved set of island-located genes.

However, one island contains virtually none of these candidate HL-defining genes. That island, ISL4, was recognized early on to contain a large concentration of cell surface-modifying genes such as enzymes for lipopolysaccharide synthesis (Coleman et al., 2006). This is especially intriguing as it suggests a scenario in which a few islands accumulated genes related to light resistance, then became fixed throughout the HL lineage. Then, a new island arose and provided an advantage in phage or predator avoidance, thus spreading through a subset of the HL clade.

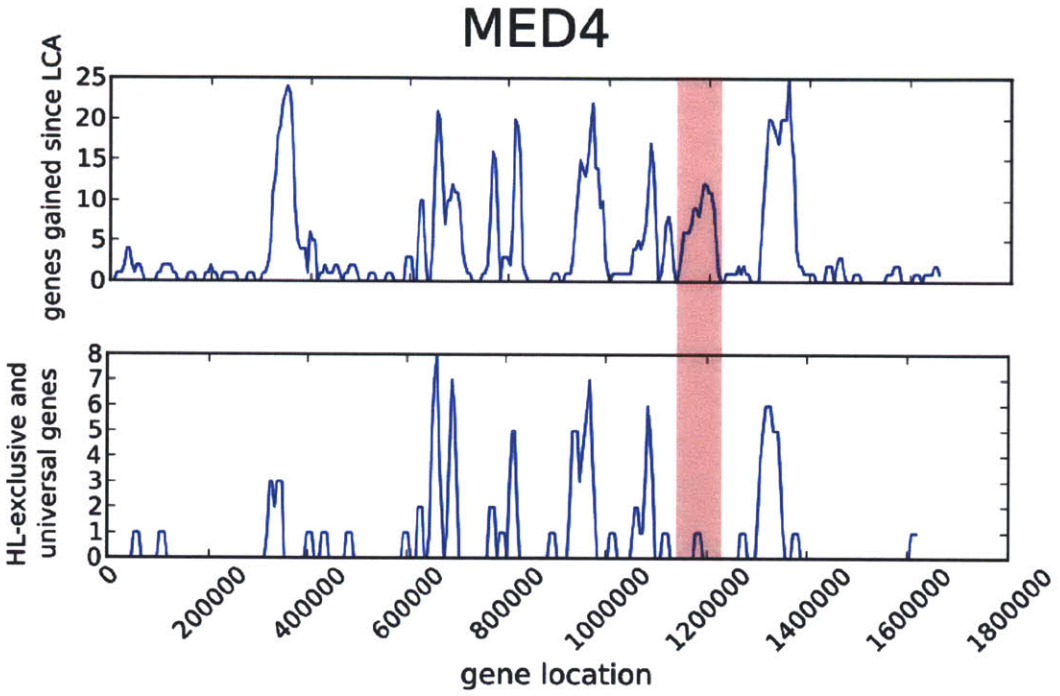


Figure I-1: Locations of possible HL-defining genes relative to MED4 islands. (A) Locations of acquired genes in MED4 since the *Prochlorococcus* last common ancestor, reproduced for reference from (Kettler et al., 2007, Chapter 2). Large peaks represent islands. (B) Locations of the 99 genes common to all HL genomes but absent from all LL genomes. On both plots, the shaded area represents ISL4, an island rich in putative cell surface-defining genes (Coleman et al., 2006).

References

Adir, N, H Zer, S Shochat, and I Ohad. Photoinhibition – a historical perspective. *Photosynthesis Research*, 76:343–370, 2003. 10.1023/A:1024969518145.

Agustí, S and M Llabrés. Solar radiation-induced mortality of marine pico-phytoplankton in the oligotrophic ocean. *Photochemistry and Photobiology*, 83(4):793–801, 2007. PMID: 17645649.

Ahlgren, N A, G Rocap, and S W Chisholm. Measurement of *Prochlorococcus* ecotypes using real-time polymerase chain reaction reveals different abundances of genotypes with similar light physiologies. *Environmental Microbiology*, 8(3):441–454, 2006.

Andersson, U, M Heddad, and I Adamska. Light stress-induced one-helix protein of the chlorophyll a/b-binding family associated with photosystem I. *Plant Physiology*, 132(2):811–820, 2003. PMID: 12805611.

Axmann, I M, U Duhring, L Seeliger, A Arnold, J T Vanselow, A Kramer, and A Wilde. Biochemical evidence for a timing mechanism in *Prochlorococcus*. *J. Bacteriol.*, 191(17):5342–5347, 2009.

Badger, J H and G J Olsen. CRITICA: coding region identification tool invoking comparative analysis. *Molecular Biology and Evolution*, 16(4):512–524, 1999. PMID: 10331277.

Bagby, S. *Life in a Drop of Water*. Ph.D. thesis, Massachusetts Institute of Technology, 2009.

Bailey, S, N H Mann, C Robinson, and D J Scanlan. The occurrence of rapidly reversible non-photochemical quenching of chlorophyll a fluorescence in cyanobacteria. *FEBS Letters*, 579(1):275–280, 2005.

Bailey, S, A Melis, K R M Mackey, P Cardol, G Finazzi, G van Dijken, G M Berg, K Arrigo, J Shrager, and A Grossman. Alternative photosynthetic electron flow to oxygen in marine *Synechococcus*. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1777(3):269–276, 2008.

Balke, V L and J D Gralla. Changes in the linking number of supercoiled DNA accompany growth transitions in *Escherichia coli*. *Journal of Bacteriology*, 169(10):4499–4506, 1987. PMID: 3308843 PMCID: 213814.

Barnett, A B. fireworx. 2007. URL `http://fireworx.sourceforge.net/`.

Bhaya, D, A Dufresne, D Vaulot, and A Grossman. Analysis of the hli gene family in marine and freshwater cyanobacteria. *FEMS Microbiology Letters*, 215(2):209–219, 2002. PMID: 12399037.

Bingham, F M and R Lukas. Seasonal cycles of temperature, salinity and dissolved oxygen observed in the Hawaii Ocean Time-series. *Deep Sea Research Part II: Topical Studies in Oceanography*, 43(2-3):199–213, 1996.

Bragg, J G, S Dutkiewicz, O Jahn, M J Follows, and S W Chisholm. Modeling selective pressures on phytoplankton in the global ocean. *PLoS ONE*, 5(3):e9569, 2010.

Campbell, L and D Vaulot. Photosynthetic picoplankton community structure in the subtropical North Pacific Ocean near Hawaii (station ALOHA). *Deep Sea Research Part I: Oceanographic Research Papers*, 40(10):2043–2060, 1993.

Choe, S E, M Boutros, A M Michelson, G M Church, and M S Halfon. Preferred analysis methods for affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biology*, 6(2):R16, 2005. PMID: 15693945.

Coleman, M L and S W Chisholm. Code and context: *Prochlorococcus* as a model for cross-scale biology. *Trends in Microbiology*, 15(9):398–407, 2007.

Coleman, M L and S W Chisholm. Ecosystem-specific selection pressures revealed through comparative population genomics. *Proceedings of the National Academy of Sciences*, 107(43):18634 –18639, 2010.

Coleman, M L, M B Sullivan, A C Martiny, C Steglich, K Barry, E F Delong, and S W Chisholm. Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science*, 311(5768):1768–1770, 2006. PMID: 16556843.

Cooper, V S and R E Lenski. The population genetics of ecological specialization in evolving *Escherichia coli* populations. *Nature*, 407(6805):736–739, 2000.

Dandonneau, Y and J Neveux. Diel variations ofin vivo fluorescence in the eastern equatorial Pacific: an unvarying pattern. *Deep Sea Research Part II: Topical Studies in Oceanography*, 44(9-10):1869–1880, 1997.

Davis, B M and M K Waldor. Filamentous phages linked to virulence of vibrio cholerae. *Current Opinion in Microbiology*, 6(1):35–42, 2003. PMID: 12615217.

Delhez, É J M and E Deleersnijder. Residence time and exposure time of sinking phytoplankton in the euphotic layer. *Journal of Theoretical Biology*, 262(3):505–516, 2010.

DeLong, E F, C M Preston, T Mincer, V Rich, S J Hallam, N Frigaard, A Martinez, M B Sullivan, R Edwards, B R Brito, S W Chisholm, and D M Karl. Community genomics among stratified microbial assemblages in the ocean's interior. *Science*, 311(5760):496 –503, 2006.

Denman, K L and A E Gargett. Time and space scales of vertical mixing and advection of phytoplankton in the upper ocean. *Limnology and Oceanography*, 28(5):801–815, 1983.

Dolganov, N A, D Bhaya, and A R Grossman. Cyanobacterial protein with similarity to the chlorophyll a/b binding proteins of higher plants: evolution and regulation. *Proceedings of the National Academy of Sciences of the United States of America*, 92(2):636–640, 1995. PMID: 7831342.

Dorman, C J. Flexible response: DNA supercoiling, transcription and bacterial adaptation to environmental stress. *Trends in Microbiology*, 4(6):214–216, 1996. PMID: 8795154.

Duce, R A and N W Tindale. Atmospheric transport of iron and its deposition in the ocean. *Limnology and Oceanography*, 36(8):1715–1726, 1991.

Dufresne, A, M Ostrowski, D J Scanlan, L Garczarek, S Mazard, B P Palenik, I T Paulsen, N T de Marsac, P Wincker, C Dossat, S Ferriera, J Johnson, A F Post, W R Hess, and F Partensky. Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. *Genome Biology*, 9(5):R90, 2008. PMID: 18507822.

Dufresne, A, M Salanoubat, F Partensky, F Artiguenave, I M Axmann, V Barbe, S Duprat, M Y Galperin, E V Koonin, F Le Gall, K S Makarova, M Ostrowski, S Oztas, C Robert, I B Rogozin, D J Scanlan, N T de Marsac, J Weissenbach, P Wincker, Y I Wolf, and W R Hess. Genome sequence of the cyanobacterium *Prochlorococcus* marinus SS120, a nearly minimal oxyphototrophic genome. *Proceedings of the National Academy of Sciences*, 100(17):10020 –10025, 2003.

Durbin, R, S Eddy, A Krogh, and G Mitchison. *Biological Sequence Analysis*. Cambridge University Press, 1998.

Edelman, M and A K Mattoo. D1-protein dynamics in photosystem II: the lingering enigma. *Photosynthesis Research*, 98(1-3):609–620, 2008.

Edgar, R C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792 –1797, 2004.

Falkowski, P G and J A Raven. *Aquatic Photosynthesis*. Princeton University Press, 2007.

Finn, R, J Clements, S Eddy, and E Rivas. Hmmer3: a new generation of sequence homology search software. 2010. URL `http://hmmer.org/`.

Frias-Lopez, J, Y Shi, G W Tyson, M L Coleman, S C Schuster, S W Chisholm, and E F DeLong. Microbial community gene expression in ocean surface waters. *Proceedings of the National Academy of Sciences*, 105(10):3805 –3810, 2008.

Frias-Lopez, J, A Thompson, J Waldbauer, and S W Chisholm. Use of stable isotope-labelled cells to identify active grazers of picocyanobacteria in ocean surface waters. *Environmental Microbiology*, 11(2):512–525, 2009. PMID: 19196281.

Funk, C and W Vermaas. A cyanobacterial gene family coding for single-helix proteins resembling part of the light-harvesting proteins from higher plants. *Biochemistry*, 38(29):9397–9404, 1999. PMID: 10413515.

Gautier, L, L Cope, B M Bolstad, and R A Irizarry. affy–analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3):307–315, 2004.

Goericke, R and N A Welschmeyer. The marine prochlorophyte *Prochlorococcus* contributes significantly to phytoplankton biomass and primary production in the Sargasso Sea. *Deep Sea Research Part I: Oceanographic Research Papers*, 40(11-12):2283–2294, 1993.

Gorbunov, M Y and P G Falkowski. Fluorescence induction and relaxation (FIRe) technique and instrumentation for monitoring photosynthetic processes and primary production in aquatic ecosystems. In *Photosynthesis: Fundamental Aspects to Global Perspectives*, pages 1029–1031. Allen Press, Montreal, 2004.

Guindon, S, J Dufayard, V Lefort, M Anisimova, W Hordijk, and O Gascuel. New algorithms and methods to estimate Maximum-Likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, 59(3):307 –321, 2010.

Hacker, J and E Carniel. Ecological fitness, genomic islands and bacterial pathogenicity. a darwinian view of the evolution of microbes. *EMBO Reports*, 2(5):376–381, 2001. PMID: 11375927.

Hakala, M, I Tuominen, M Keränen, T Tyystjärvi, and E Tyystjärvi. Evidence for the role of the oxygen-evolving manganese complex in photoinhibition of photosystem II. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1706(1-2):68–80, 2005.

Havaux, M, G Guedeney, Q He, and A R Grossman. Elimination of high-light-inducible polypeptides related to eukaryotic chlorophyll a/b-binding proteins results in aberrant photoacclimation in *Synechocystis* PCC6803. *Biochimica Et Biophysica Acta*, 1557(1-3):21–33, 2003. PMID: 12615345.

He, Q, N Dolganov, O Bjorkman, and A R Grossman. The high light-inducible polypeptides in *Synechocystis* PCC6803. expression and function in high light. *The Journal of Biological Chemistry*, 276(1):306–314, 2001. PMID: 11024039.

Huner, N P A, G Öquist, and F Sarhan. Energy balance and acclimation to light and cold. *Trends in Plant Science*, 3(6):224–230, 1998.

Jantaro, S, Q Ali, S Lone, and Q He. Suppression of the lethality of high light to a quadruple HLI mutant by the inactivation of the regulatory protein PfsR in *Synechocystis* PCC 6803. *Journal of Biological Chemistry*, 281(41):30865 –30874, 2006.

Jeong, K S, J Ahn, and A B Khodursky. Spatial patterns of transcriptional activity in the chromosome of *Escherichia coli*. *Genome Biology*, 5(11):R86, 2004. PMID: 15535862.

Johnson, Z I, E R Zinser, A Coe, N P McNulty, E M S Woodward, and S W Chisholm. Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science*, 311(5768):1737–1740, 2006. PMID: 16556835.

Jürgens, K and C Matz. Predation as a shaping force for the phenotypic and genotypic composition of planktonic bacteria. *Antonie van Leeuwenhoek*, 81:413–434, 2002.

Karl, D M and R Lukas. The Hawaii Ocean Time-series (HOT) program: Background, rationale and field implementation. *Deep Sea Research Part II: Topical Studies in Oceanography*, 43(2-3):129–156, 1996.

Kettler, G C, A C Martiny, K Huang, J Zucker, M L Coleman, S Rodrigue, F Chen, A Lapidus, S Ferriera, J Johnson, C Steglich, G M Church, P Richardson, and S W Chisholm. Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genetics*, 3(12):e231, 2007.

Kiefer, D A. Chlorophyll a fluorescence in marine centric diatoms: Responses of chloroplasts to light and nutrient stress. *Marine Biology*, 23(1):39–46, 1973.

Kolber, Z S, O Prásil, and P G Falkowski. Measurements of variable chlorophyll fluorescence using fast repetition rate techniques: defining methodology and experimental protocols. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1367(1-3):88–106, 1998.

Kufryk, G, M A Hernandez-Prieto, T Kieselbach, H Miranda, W Vermaas, and C Funk. Association of small CAB-like proteins (SCPs) of *Synechocystis* sp. PCC 6803 with photosystem II. *Photosynthesis Research*, 95(2-3):135–145, 2008. PMID: 17912610.

Lande, R and A M Wood. Suspension times of particles in the upper ocean. *Deep Sea Research Part A. Oceanographic Research Papers*, 34(1):61–72, 1987.

Letunic, I and P Bork. Interactive tree of life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23(1):127 –128, 2007.

Lindell, D, J D Jaffe, M L Coleman, M E Futschik, I M Axmann, T Rector, G Kettler, M B Sullivan, R Steen, W R Hess, G M Church, and S W Chisholm. Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. *Nature*, 449(7158):83–86, 2007.

Lindell, D, J D Jaffe, Z I Johnson, G M Church, and S W Chisholm. Photosynthesis genes in marine viruses yield proteins during host infection. *Nature*, 438(7064):86–89, 2005.

Lindell, D, M B Sullivan, Z I Johnson, A C Tolonen, F Rohwer, and S W Chisholm. Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proceedings of the National Academy of Sciences of the United States of America*, 101(30):11013–11018, 2004. PMID: 15256601.

Lipschultz, F, N R Bates, C A Carlson, and D A Hansell. New production in the Sargasso Sea: History and current status. *Global Biogeochemical Cycles*, 16(1), 2002.

Llabrés, M and S Agustí. Picophytoplankton cell death induced by UV radiation: Evidence for oceanic Atlantic communities. *Limnology and Oceanography*, 51(1):21–29, 2006.

Loftus, M E and H H Seliger. Some limitations of the in vivo fluorescence technique. *Chesapeake Science*, 16(2):79, 1975.

Long, S P, S Humphries, and P G Falkowski. Photoinhibition of photosynthesis in nature. *Annual Review of Plant Physiology and Plant Molecular Biology*, 45(1):633–662, 1994.

Macintyre, H L and J J Cullen. Using cultures to investigate the physiological ecology of microalgae. In Robert A. Anderson, editor, *Algal Culturing Techniques*, pages 287–326. Elsevier Academic Press, Burlington, 2005.

Mackey, K R M, A Paytan, A R Grossman, and S Bailey. A photosynthetic strategy for coping in a high-light, low-nutrient environment. *Limnology and Oceanography*, 53(3):900–913, 2008.

Malmstrom, R R, A Coe, G C Kettler, A C Martiny, J Frias-Lopez, E R Zinser, and S W Chisholm. Temporal dynamics of *Prochlorococcus* ecotypes in the Atlantic and Pacific oceans. *ISME J*, 4(10):1252–1264, 2010.

Marioni, J, C Mason, S Mane, M Stephens, and Y Gilad. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 2008. PMID: 18550803.

Martiny, A C, M L Coleman, and S W Chisholm. Phosphate acquisition genes in *Prochlorococcus* ecotypes: Evidence for genome-wide adaptation. *Proceedings of the National Academy of Sciences*, 103(33):12552 –12557, 2006.

Martiny, A C, Y Huang, and W Li. Occurrence of phosphate acquisition genes in *Prochlorococcus* cells from different ocean regions. *Environmental Microbiology*, 11(6):1340–1347, 2009a. PMID: 19187282.

Martiny, A C, A P K Tai, D Veneziano, F Primeau, and S W Chisholm. Taxonomic resolution, ecotypes and the biogeography of *Prochlorococcus*. *Environmental Microbiology*, 11(4):823–832, 2009b.

Mary, I, C Tu, A Grossman, and D Vaulot. Effects of high light on transcripts of stress-associated genes for the cyanobacteria *Synechocystis* sp. PCC 6803 and *Prochlorococcus* MED4 and MIT9313. *Microbiology*, 150(5):1271–1281, 2004.

Mary, I and D Vaulot. Two-component systems in *Prochlorococcus* MED4: genomic analysis and differential expression under stress. *FEMS Microbiology Letters*, 226(1):135–144, 2003.

Melis, A. Photosystem-II damage and repair cycle in chloroplasts: what modulates the rate of photodamage in vivo? *Trends in Plant Science*, 4(4):130–135, 1999.

Mirold, S, W Rabsch, M Rohde, S Stender, H Tschäpe, H Rüssmann, E Igwe, and W D Hardt. Isolation of a temperate bacteriophage encoding the type III effector protein SopE from an epidemic *Salmonella typhimurium* strain. *Proceedings of the National Academy of Sciences of the United States of America*, 96(17):9845–9850, 1999. PMID: 10449782.

Moore, L R and S W Chisholm. Photophysiology of the marine cyanobacterium *Prochlorococcus* : ecotypic differences among cultured isolates. *Limnology and Oceanography*, 44(3):628–638, 1999.

Moore, L R, A Coe, E R Zinser, M A Saito, M B Sullivan, D Lindell, K Frois-Moniz, J Waterbury, and S W Chisholm. Culturing the marine cyanobacterium *Prochlorococcus*. *Limnology and Oceanography: Methods*, 5:353–362, 2007.

Moore, L R, G Rocap, and S W Chisholm. Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature*, 393(6684):464–467, 1998.

Morris, J J, R Kirkegaard, M J Szul, Z I Johnson, and E R Zinser. Robust growth of *Prochlorococcus* colonies and dilute liquid cultures: facilitation by "helper" heterotrophic bacteria. *Appl. Environ. Microbiol.*, pages AEM.02479–07, 2008.

Morris, J. Jeffrey, Zackary I. Johnson, Martin J. Szul, Martin Keller, and Erik R. Zinser. Dependence of the cyanobacterium *Prochlorococcus* on hydrogen peroxide scavenging microbes for growth at the ocean's surface. *PLoS ONE*, 6(2):e16805, 2011.

Nishiyama, Y, S I Allakhverdiev, and N Murata. A new paradigm for the action of reactive oxygen species in the photoinhibition of photosystem II. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1757(7):742–749, 2006.

Nixon, P J, F Michoux, J Yu, M Boehm, and J Komenda. Recent advances in understanding the assembly and repair of photosystem II. *Ann Bot*, page mcq059, 2010.

Notredame, C, D G Higgins, and J Heringa. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302(1):205–217, 2000. PMID: 10964570.

Osburne, M S, B M Holmbeck, J Frias-Lopez, R Steen, K Huang, L Kelly, A Coe, K Waraska, A Gagne, and S W Chisholm. UV hyper-resistance in *Prochlorococcus* MED4 results from a single base pair deletion just upstream of an operon encoding nudix hydrolase and photolyase. *Environmental Microbiology*, 12(7):1978–1988, 2010. PMID: 20345942.

Pace, N R. A molecular view of microbial diversity and the biosphere. *Science*, 276(5313):734 –740, 1997.

Pan, X, M Li, T Wan, L Wang, C Jia, Z Hou, X Zhao, J Zhang, and W Chang. Structural insights into energy regulation of light-harvesting complex CP29 from spinach. *Nat Struct Mol Biol*, 18(3):309–315, 2011.

Partensky, F, W R Hess, and D Vaulot. *Prochlorococcus* , a marine photosynthetic prokaryote of global significance. *Microbiol. Mol. Biol. Rev.*, 63(1):106–127, 1999.

Peter, B J, J Arsuaga, A M Breier, A B Khodursky, P O Brown, and N R Cozzarelli. Genomic transcriptional response to loss of chromosomal supercoiling in *Escherichia coli*. *Genome Biology*, 5(11):R87, 2004. PMID: 15535863.

Promnares, K, J Komenda, L Bumba, J Nebesarova, F Vacha, and M Tichy. Cyanobacterial small chlorophyll-binding protein ScpD (HliB) is located on the periphery of photosystem II in the vicinity of PsbH and CP47 subunits. *J. Biol. Chem.*, 281(43):32705–32713, 2006.

Rappe, M S, S A Connon, K L Vergin, and S J Giovannoni. Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature*, 418(6898):630–633, 2002.

Rocap, G, D L Distel, J B Waterbury, and S W Chisholm. Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. *Appl. Environ. Microbiol.*, 68(3):1180–1191, 2002.

Rocap, G, F W Larimer, J Lamerdin, S Malfatti, P Chain, N A Ahlgren, A Arellano, M Coleman, L Hauser, W R Hess, Z I Johnson, M Land, D Lindell, A F Post, W Regala, M Shah, S L Shaw, C Steglich, M B Sullivan, C S Ting, A Tolonen, E A Webb, E R Zinser, and S W Chisholm. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature*, 424(6952):1042–1047, 2003.

Rodrigue, S, R R Malmstrom, A M Berlin, B W Birren, M R Henn, and S W Chisholm. Whole genome amplification and de novo assembly of single bacterial cells. *PLoS ONE*, 4(9):e6864, 2009.

Rodrigue, S, A C Materna, S C Timberlake, M C Blackburn, R R Malmstrom, E J Alm, and S W Chisholm. Unlocking short read sequencing for metagenomics. *PLoS ONE*, 5(7), 2010.

Rusch, D B, A C Martiny, C L Dupont, A L Halpern, and J C Venter. Characterization of *Prochlorococcus* clades from iron-depleted oceanic regions. *Proceedings of the National Academy of Sciences*, 107(37):16184 –16189, 2010.

Rusch, D B et al. The sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biology*, 5(3):e77, 2007. PMID: 17355176.

Röttgers, R. Comparison of different variable chlorophyll a fluorescence techniques to determine photosynthetic parameters of natural phytoplankton. *Deep Sea Research Part I: Oceanographic Research Papers*, 54(3):437–451, 2007.

Sher, D, J W Thompson, N Kashtan, L Croal, and S W Chisholm. Response of *Prochlorococcus* ecotypes to co-culture with diverse marine bacteria. *The ISME Journal*, 2011. PMID: 21326334.

Six, C, Z V Finkel, A J Irwin, and D A Campbell. Light variability illuminates niche-partitioning among marine picocyanobacteria. *PLoS ONE*, 2(12):e1341, 2007.

Steglich, C, M Futschik, T Rector, R Steen, and S W Chisholm. Genome-wide analysis of light sensing in *Prochlorococcus*. *Journal of Bacteriology*, 188(22):7796–7806, 2006. PMID: 16980454.

Steglich, C, D Lindell, M Futschik, T Rector, R Steen, and S W Chisholm. Short RNA half-lives in the slow-growing marine cyanobacterium *Prochlorococcus*. *Genome Biology*, 11(5):R54, 2010.

Steinberg, D K, C A Carlson, N R Bates, R J Johnson, A F Michaels, and A H Knap. Overview of the US JGOFS Bermuda Atlantic Time-series Study (BATS): a decade-scale look at ocean biology and biogeochemistry. *Deep Sea Research Part II: Topical Studies in Oceanography*, 48(8-9):1405–1447, 2001.

Storm, P, M A Hernandez-Prieto, L L Eggink, J K Hoober, and C Funk. The small CAB-like proteins of *Synechocystis* sp. PCC 6803 bind chlorophyll. *Photosynthesis Research*, 98(1-3):479–488, 2008.

Sullivan, M B, M L Coleman, P Weigele, F Rohwer, and S W Chisholm. Three *Prochlorococcus* cyanophage genomes: Signature features and ecological interpretations. *PLoS Biology*, 3(5), 2005. PMID: 15828858 PMCID: 1079782.

Sullivan, M B, J B Waterbury, and S W Chisholm. Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*. *Nature*, 424(6952):1047–1051, 2003.

Thompson, A. *Iron and* Prochlorococcus. Ph.D. thesis, Massachusetts Institute of Technology, 2009.

Thompson, A W, K Huang, M A Saito, and S W Chisholm. Transcriptome response of high- and low-light adapted *Prochlorococcus* strains to changing iron availability. *ISME J*, 2011.

Thompson, L R, Q Zeng, L Kelly, K H Huang, M L Coleman, A U Singer, J Stubbe, and S W Chisholm. Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proceedings of the National Academy of Sciences*, In press.

Tolonen, A C, J Aach, D Lindell, Z I Johnson, T Rector, R Steen, G M Church, and S W Chisholm. Global gene expression of *Prochlorococcus* ecotypes in response to changes in nitrogen availability. *Molecular Systems Biology*, 2:53, 2006. PMID: 17016519.

Tringe, S G, C von Mering, A Kobayashi, A A Salamov, K Chen, H W Chang, M Podar, J M Short, E J Mathur, J C Detter, P Bork, P Hugenholtz, and E M Rubin. Comparative metagenomics of microbial communities. *Science*, 308(5721):554 –557, 2005.

Tyson, G W, J Chapman, P Hugenholtz, E E Allen, R J Ram, P M Richardson, V V Solovyev, E M Rubin, D S Rokhsar, and J F Banfield. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43, 2004. PMID: 14961025.

Ueno, H and T Yonesaki. Phage-induced change in the stability of mRNAs. *Virology*, 329(1):134–141, 2004.

Vavilin, D, D Yao, and W Vermaas. Small cab-like proteins retard degradation of photosystem II-associated chlorophyll in *Synechocystis* sp. PCC 6803. *Journal of Biological Chemistry*, 282(52):37660–37668, 2007.

Wai, S N, K Nakayama, K Umene, T Moriya, and K Amako. Construction of a ferritin-deficient mutant of Campylobacter jejuni: contribution of ferritin to iron storage and protection against oxidative stress. *Molecular Microbiology*, 20(6):1127–1134, 1996. PMID: 8809765.

Wang, Q, S Jantaro, B Lu, W Majeed, M Bailey, and Q He. The high light-inducible polypeptides stabilize trimeric photosystem I complex under high light conditions in *Synechocystis* PCC 6803. *Plant Physiology*, 147(3):1239–1250, 2008. PMID: 18502976.

Waterbury, J B and F W Valois. Resistance to Co-Occurring phages enables marine *Synechococcus* communities to coexist with cyanophages abundant in seawater. *Appl. Environ. Microbiol.*, 59(10):3393–3399, 1993.

West, N J and D J Scanlan. Niche-partitioning of *Prochlorococcus* populations in a stratified water column in the Eastern North Atlantic Ocean. *Appl. Environ. Microbiol.*, 65(6):2585–2591, 1999.

Willenbrock, H and D W Ussery. Chromatin architecture and gene expression in *Escherichia coli*. *Genome Biology*, 5(12):252, 2004. PMID: 15575978.

Wilmes, P, S L Simmons, V J Denef, and J F Banfield. The dynamic genetic repertoire of microbial communities. *FEMS Microbial Rev*, 33:109–132, 2008.

Wu, J, W Sunda, E A Boyle, and D M Karl. Phosphate depletion in the Western North Atlantic Ocean. *Science*, 289(5480):759 –762, 2000.

Yao, D, T Kieselbach, J Komenda, K Promnares, M A H Prieto, M Tichy, W Vermaas, and C Funk. Localization of the small CAB-like proteins in photosystem II. *The Journal of Biological Chemistry*, 282(1):267–276, 2007. PMID: 17105726.

Ziegelhoffer, E C and T J Donohue. Bacterial responses to photo-oxidative stress. *Nat Rev Micro*, 7(12):856–863, 2009.

Zinser, E, Z I Johnson, A Coe, E Karaca, D Veneziano, and S W Chisholm. Influence of light and temperature on *Prochlorococcus* ecotype distributions in the Atlantic Ocean. *Limnology and Oceanography*, 52(5):2205–2220, 2007.

Zinser, E R, A Coe, Z I Johnson, A C Martiny, N J Fuller, D J Scanlan, and S W Chisholm. *Prochlorococcus* ecotype abundances in the North Atlantic Ocean as revealed by an improved quantitative PCR method. *Appl. Environ. Microbiol.*, 72(1):723–732, 2006.

Zinser, E R, D Lindell, Z I Johnson, M E Futschik, C Steglich, M L Coleman, M A Wright, T Rector, R Steen, N McNulty, L R Thompson, and S W Chisholm. Choreography of the transcriptome, photophysiology, and cell cycle of a minimal photoautotroph, *Prochlorococcus*. *PLoS ONE*, 4(4):e5135, 2009. PMID: 19352512.